UNIVERSIDAD
NACIONAL
DE COLOMBIA

En búsqueda de consenso sobre el concepto de validez: Un estudio Delphi

Sandra Liliana Camargo Salamanca

Universidad Nacional de Colombia

Facultad de Ciencias Humanas

Departamento de Psicología

2017

En búsqueda de consenso sobre el concepto de validez: Un estudio Delphi

Sandra Liliana Camargo Salamanca

Tesis para obtener el título de Magíster en Psicología con énfasis en investigación

Dirigido por:

Aura Nidia Herrera Rojas (Ph.D.)

Codirigido por:

Anne Traynor (Ph.D.)

Línea de investigación:

Métodos e instrumentos para la investigación en ciencias del comportamiento

Universidad Nacional de Colombia

Facultad de Ciencias Humanas

Departamento de Psicología

2017

Dedicatoria

A mis tías Eugenia, Janec y Yomara,
a mi tío Uriel,
y a mi mamá,
por todo el amor que me han dado y
por su apoyo para poder alcanzar las metas que me he propuesto.

Agradecimientos

A los expertos, Denny Borsboom, Gregory Cizek, Michael Kane, Robert Mislevy, José Muñiz, Paul Newton y Stuart Shaw por su tiempo, compromiso y aportes durante el desarrollo de esta investigación.

A la profesora Aura Nidia Herrera Rojas por sus orientaciones y acompañamiento en el proceso de investigación.

A la profesora Anne Traynor por su valiosa asesoría a lo largo del desarrollo del estudio Delphi.

A todo el equipo asesor del Delphi, Rocío Barajas, Víctor H. Cervantes y Jazmine Escobar, por sus valiosos aportes, su compromiso y su acompañamiento en todo el estudio.

A los integrantes del seminario de investigación *Métodos e instrumentos para la investigación en ciencias del comportamiento* por sus aportes.

A toda mi familia por el apoyo emocional y económico que me prestaron para que se pudiera realizar este trabajo.

Resumen

El objetivo de este trabajo es identificar, entre un grupo de expertos, los puntos en los que existe consenso y aquellos en los que no, sobre el concepto de validez en educación y psicología, por medio de un estudio Delphi en línea, con el fin de aportar elementos que contribuyan a dar claridad y consistencia a la conceptualización de la validez. Es necesario explorar las diferentes opiniones sobre estos temas e identificar los temas en los que se presenta acuerdo y aquellos en los que no debido a que a) se han identificado diferentes posturas en la conceptualización de la *validez,* algunas muy diferentes e incluso encontradas; y a que, b) si bien los expertos han expresado sus opiniones sobre este concepto, existen temas en los que solo algunos expertos han opinado y falta recoger las ideas de otros expertos al respecto. Los participantes fueron expertos académicos reconocidos que han liderado la discusión sobre el concepto de *validez* en las últimas décadas tanto en publicaciones de alto nivel como en reuniones académicas en Europa y en Estados Unidos. Los resultados del estudio Delphi muestran cuatro categorías de análisis: concepto de validez, Estándares (2014), consenso y validación. Se identificaron algunos puntos muertos de la discusión sobre el concepto de validez y se identificaron las ideas en las que se halló consenso y recomendaciones para cada una de las categorías de análisis, las cuales contribuyen en el avance de la consolidación del concepto de validez.

Palabras clave: Concepto de validez, estudio Delphi, consenso.

Abstract

The purpose of this work is to identify, among a group of experts, the issues in which there is consensus and those in which there is not regarding the concept of validity in educational and psychological testing, by means of an online Delphi study, and thus contribute to give clarity and consistency to the conceptualization of validity. The opinions of these experts should be explored because different conceptual positions have been identified when approaching validity, some of them very different and even incompatible. Also, although several experts have expressed their views on this concept, there are issues in which only a few of them have expressed their opinion and it is important to collect ideas from other experts. Participants in this study were recognized academic experts who have led the discussion on the concept of validity in recent decades in both high-level publications and academic meetings in Europe and the United States. The Delphi study's results show four categories: the concept of validity, the Standards (2014), consensus, and validation. Some of the ideas around the concept of validity for which there is a standoff among the experts' opinions as well as issues for each category for which a consensus could be reached were identified. Some recommendations to advance the conceptualization of validity are reached for each category.

Keywords: Concept of validity, Delphi study, consensus.

Tabla de contenido

Lista de tablas

Lista de figuras

Introducción

La evaluación en la educación y la psicología ha recurrido al uso de pruebas como una de las principales fuentes de recolección de información a nivel grupal e individual. El diseño, el uso y la interpretación apropiadas de las pruebas buscan orientar la toma de decisiones respecto a individuos, grupos o programas evaluados, y brindar solidez a la producción teórica en psicología, siendo esta última condición básica para la consolidación de la Psicología como ciencia. Por lo cual, la creación, el uso y el estudio de métodos e instrumentos válidos, que permitan llegar a conclusiones cada vez más robustas, son necesidades básicas del campo investigativo de la psicología (AERA, APA y NCME, 2014).

Específicamente, dentro del estudio de los instrumentos de evaluación, tanto a nivel teórico como metodológico, se presentan temas vitales de investigación y desarrollo como los estándares de calidad de los instrumentos, los modelos de respuesta, la confiabilidad y la validez. Los estudios teóricos de estos temas generan avance de la psicología como ciencia y permiten reflexionar sobre la contribución de la psicología al mundo científico y social. Uno de estos temas más discutidos en la última década en congresos, publicaciones seriadas y en textos teóricos de metodología de investigación y psicometría es la *validez*. Las discusiones teóricas relacionadas con este concepto han enmarcado posturas que orientan metodologías para evaluar comportamientos y han delimitado el alcance de las conclusiones que se obtienen con instrumentos como las pruebas objetivas o test.

En varias ocasiones, se han presentado periodos de discusión alrededor de la validez en el contexto de la medición en psicología. Las discusiones iniciales, durante los años veinte y treinta del siglo XX, giraron alrededor del aspecto metodológico, particularmente, se centraron en establecer cómo se podría mostrar que una prueba psicológica fuera válida; de esta discusión surgieron una variedad de procedimientos para evaluar la validez. Durante la década de los cincuentas, la discusión se tornó hacia el problema de la definición conceptual de qué es validez; como resultado de esta discusión, se llegó a la definición de un conjunto de tipos de validez que organizaron los procedimientos surgidos anteriormente (Lissitz, 2009).

En 1955, Cronbach y Meehl plantearon esta discusión por primera vez en el artículo «*Concept of Validity*», el cual ubicó el tema de los constructos psicológicos y del uso e interpretación de los resultados de las pruebas en el centro de la conceptualización de la validez. La edición de los «Standards for Educational and Psychological Testing» (en adelante, los Estándares) de la AERA, la APA y el NCME de 1985 disparó nuevamente la discusión sobre el concepto de validez al afirmar que esta es «*la consideración más importante en la valoración de una prueba*». Como resultado de este periodo surgió una definición de *validez* que buscó resaltar la idea de que esta se debía conceptualizar independientemente de las formas con que se recogiera información o evidencias de ella y buscó darle unidad a dicho concepto. Esta propuesta fue presentada por Messick (1988, 1989) y guió la llamada visión unificada del concepto de validez que se incorporó en la edición de los Estándares de la AERA, la APA y el NCME de 1999.

La discusión actual, que se ha venido dando desde la última década del siglo pasado y la primera del siglo XXI, ha presentado una gran variedad de críticas y defensas a la visión unificada de *validez* y se ha dirigido hacia la búsqueda de una conceptualización más sólida de los distintos términos empleados en su definición. Muchos de los autores que estudian actualmente la *validez* están de acuerdo en que el problema principal es la falta de claridad en la definición del concepto (Lissitz, 2009; Newton, 2013a, 2013b); sin embargo, no existe acuerdo sobre cómo este se puede aclarar. Lissitz (2009) identifica dos corrientes diferentes con las que se busca conseguir este fin: una que parte de la postura oficial de la AERA, la APA y el NCME (1999) y se apoya en la conceptualización de validez unificada de Messick (1989), y otra que insiste en la necesidad de una reconceptualización de la validez, ya que cuestiona principios básicos de la propuesta de validez unificada.

Otro de los aspectos en los que se centra esta discusión es la conceptualización de validez que opera en la puesta en práctica de los procesos de validación, propuesta por Moss (2007) y Sireci (2007). De acuerdo con Lissitz (2009), estos autores proponen que se incluyan, en la edición de los Estándares, ejemplos de cómo se pueden conseguir evidencias de validez en casos concretos de procesos de validación, además de los requerimientos de validez. En este sentido, existe un cierto acuerdo en que la literatura sobre *validez* no ha abordado el problema de la puesta en práctica del concepto. Por su

parte, los métodos estadísticos y la metodología de investigación que se usan o que pueden usarse para la investigación de la validez de una prueba no han sido parte central de la discusión ni tienen un lugar particular dentro de los Estándares de 1999. Un último aspecto central en esta discusión proviene de la postura adoptada con respecto a lo que se entiende por medición. Al respecto, Markus y Borsboom (2013) identifican cuatro posturas filosóficas, cada una de las cuales conlleva una manera distinta de aproximarse a la medición que tiene efectos sobre cómo entender la *validez* y sobre cómo se debe realizar un proceso de validación.

Teniendo en cuenta este panorama, expertos en el estudio de la *validez* consideran que en este campo específico de la psicometría se encuentran dos necesidades primordiales. Por una parte, hay una necesidad de crear teoría sobre validez, la cual, de acuerdo con la corriente desde la que se aborde el problema, puede estar más cercana a una necesidad de aclarar los conceptos relacionados con la *validez* (p. ej. Hubley y Zumbo, 2011) o con una redefinición del concepto mismo (p. ej. Borsboom, Mellenbergh y Van Heerden, 2004; Lissitz y Samuelsen, 2007; Cizek, 2012); y, por otra parte, hay una necesidad de aclarar la relación entre el concepto de validez y el proceso de validación, y los procedimientos utilizados para llevar a cabo este último. Con relación a la primera necesidad, es importante reconocer que para poder avanzar en la teoría se requiere comprender el estado de la discusión en el momento actual, por lo tanto, en este trabajo se busca analizar las diversas posturas propuestas en la discusión actual sobre el concepto de validez; y, con relación a la segunda necesidad, debe tenerse en cuenta que la definición de estas relaciones requiere asumir una postura sobre la primera necesidad, por lo cual, desarrollar este aspecto sobrepasa los objetivos de este trabajo.

Algunos de los estudios que se han hecho a nivel nacional orientados a satisfacer estas necesidades, particularmente a identificar fuentes de invalidez, han sido desarrollados, en la última década, por el grupo *Métodos e instrumentos para la investigación en ciencias del comportamiento* del Departamento de Psicología de la Universidad Nacional de Colombia, y han tenido como objetivo identificar fuentes de invalidez de las pruebas que pueden afectar la equidad de la calificación, con énfasis en los instrumentos desarrollados por el Instituto Colombiano para la Evaluación de la Educación (Icfes), entidad que realiza y

administra los Exámenes de Estado, uno de estos es Saber 11.º, cuyos resultados son utilizados por las Instituciones de Educación Superior para seleccionar a sus estudiantes con base en un puntaje mínimo definido por ellas mismas (Congreso de la República de Colombia, 2009), por lo tanto, los resultados de estas evaluaciones tienen un alto impacto en la vida de evaluados (Cuevas, 2013).

Los primeros estudios desarrollados por el grupo relacionados con las fuentes de invalidez en las pruebas se centraron en la investigación de procedimientos para la detección de Funcionamiento Diferencial de los Ítems (DIF, por sus siglas en inglés) (Herrera, 2005; Herrera, Gómez y Hidalgo, 2005; Herrera, Gómez y Muñiz, 2007; Herrera, Gómez, Quintero, Arias, Berrío y Cervantes, 2007; Berrío, 2008; Arias, 2008; Santana, 2009); luego, algunos trabajos se enfocaron en el estudio de metodologías para la identificación de las posibles causas del DIF, es decir, análisis de sesgo (Cuevas, 2013; Rico, 2015) y otras investigaciones se orientaron hacia estudios de equiparación y al desarrollo de bancos de preguntas y Test Adaptativos Informatizados (TAI) sobre comprensión lectora para personas con y sin limitación visual (Lancheros, 2013; Espinosa, 2014; Soler, 2014; Casas, 2016; Rodríguez, 2016; Barajas, 2017). El interés del grupo por el estudio de la validez ha aportado avances a nivel teórico de la psicometría y a nivel práctico-social en nuestro país. Siguiendo esta línea de investigación, el presente proyecto se orientó a estudiar la validez de las pruebas con un enfoque más teórico que metodológico, el cual complementa el trabajo realizado por el grupo en sus investigaciones anteriores.

Objetivo general

El objetivo de este trabajo es identificar, entre un grupo de expertos, los puntos en los que existe consenso y aquellos en los que no sobre el concepto de validez en educación y psicología, por medio del desarrollo de un estudio Delphi en línea, con el fin de aportar elementos que contribuyan a dar claridad y consistencia a la conceptualización de la validez.

Objetivos específicos

Para alcanzar el logro del objetivo general se proponen los siguientes objetivos específicos: a) identificar las posturas teóricas, históricas y actuales, más relevantes que han orientado la discusión sobre el concepto de validez, b) identificar los principales cuestionamientos actuales que abordan el desarrollo de la teoría de la validez y las posturas de expertos en cada uno de ellos; y finalmente, c) identificar los elementos en los que ellos llegan a un consenso y en los que no, respecto a las opiniones expuestas por parte de expertos, sobre al concepto de validez por medio del estudio Delphi.

Revisión bibliográfica

Perspectivas teóricas del estudio de la validez

A continuación, se presenta el desarrollo de las diversas posturas en la discusión sobre el concepto de validez expuestas por expertos del área de medición y evaluación del comportamiento. La presentación de la discusión del concepto de validez se dividió en cuatro partes: En la primera, se tiene en cuenta la orientación pragmática desarrollada desde los años diez hasta los cincuenta; en la segunda, se presenta la postura unificada de la validez; en la tercera, se exponen las posturas actuales sobre la discusión del concepto de validez; y por último, la viabilidad de un consenso sobre la definición de validez. Adicionalmente, en este capítulo se presenta la estructura del Método Delphi, el cual fue el procedimiento escogido para realizar el análisis de las opiniones de los expertos en la búsqueda de alcanzar un consenso sobre el concepto de validez.

Concepto de validez desde la perspectiva pragmática

El nacimiento del concepto de validez no se puede establecer sin antes identificar el contexto de la evaluación a inicios del siglo XX. En estos años, con la creación de baterías de medición de habilidades mentales como la inteligencia, la memoria y la atención realizadas por Binet (Binet 1905; Binet y Henri, 1899), se asumió que el puntaje era un indicador de la habilidad mental, que existía una relación lineal entre este puntaje y la habilidad, y que esta habilidad era unidimensional (Kane, 2013); dichas pruebas tuvieron gran difusión y se usaron de manera masiva en el reclutamiento para la armada en la Primera Guerra Mundial.

Newton y Shaw (2014) relatan cómo la gran acogida que tuvieron los instrumentos creados por Binet (1905) motivaron la creación de otros test en el campo educativo y psicológico, y cómo la difusión de nuevas pruebas y de publicaciones, como las de Thorndike en 1903 y 1904 sobre el desarrollo de pruebas en el ámbito educativo y la teoría de la medición mental y social, promovieron el interés de muchos profesionales de las áreas de educación y psicología por el diseño, la creación de test y el análisis de sus resultados. No obstante, como consecuencia de la proliferación del uso de las pruebas, la escuela

tradicional empezó a mostrar desacuerdo con el uso de los resultados e interpretaciones que se hacían de ellas, y demostró la falta de consistencia de las evaluaciones (Meyer, 1908; Starch y Elliot, 1912, 1913), este descontento fue vital para el nacimiento del estudio de la validez y de la confiabilidad de la evaluación, y para el desarrollo de investigaciones en el área, ya que, como afirma Freeman (1917), las publicaciones con referencia a estudios experimentales sobre pruebas incrementaron más del doble en tan solo cinco años.

Este contexto promovió en las comunidades académicas y profesionales la necesidad de discutir, definir y unificar las interpretaciones de los conceptos y de los estándares de calidad para el desarrollo y la aplicación de pruebas. Buckingham, McCall, Otis, Rugg, Trabue y Courtis (1921) relatan que dos de las principales preocupaciones de los miembros del Comité de Estandarización de la Asociación Nacional de Directores de Investigación Educativa fueron la validez y la confiabilidad de las pruebas y también narran que el comité propuso resolver el problema de validez por medio de la determinación de la relación entre las puntuaciones realizadas en una prueba y otras medidas de la misma habilidad.

Las discusiones del tipo de evidencia que sustentaron las conclusiones sobre la validez de un test se dividieron entre el análisis lógico y el análisis pragmático o empírico (Rulon, 1946). El análisis lógico recoge evidencia sobre el contenido de lo que se está evaluando por medio de juicios de expertos en el tema evaluado y el análisis empírico es el intento por investigar el grado en que la prueba está correlacionada con otra prueba y, por lo tanto, miden lo mismo (Cronbach, 1949).

Newton y Shaw (2014) afirman que, si bien en las primeras discusiones se le dio gran importancia a la correlación entre la prueba y el criterio como fuente validez, esta no fue la única idea presente en esta etapa, el análisis lógico o de contenido también estuvo presente en diversas posturas orientándose a la búsqueda de evidencias y al contraste de unas con otras. Un ejemplo de estas propuestas fue el trabajo de Monroe (1923) que presentó la *validez* como el grado en que se mantiene una relación funcional entre los puntajes de la prueba y las habilidades específicas que son medidas en dicha prueba. Además de este aspecto correlacional, en su definición, Monroe (1923) también centró su atención en el contenido de lo que evalúa la prueba. En la descripción que Monroe (1923) hace del proceso para hallar evidencia de la validez toma en cuenta la evaluación de la

operacionalización de las conductas a evaluar, la confiabilidad, la discriminación, la comparación con otras medidas criterio y de las inferencias basadas en la estructura del instrumento y en su administración. Otros autores, dentro de sus discusiones sobre el concepto de validez, también estudiaron la relación del criterio de jueces contra la correlación (Kelly, 1927), la opinión de los expertos versus la experimentación (Ruch, 1929), y currículo versus estadística (Ruch, 1933); en medio estas primeras discusiones sobre la validez por parte de educadores y académicos, surgió la definición clásica de validez: «*grado en el que una prueba mide lo que se propone medir*» (Ruch, 1924, p. 13).

Sireci (2009) denota que, en sus inicios, la definición de validez fue pragmática y se estructuró, principalmente, en términos de correlación de los puntajes de la prueba con algún criterio, lo cual fue influenciado y animado por la publicación de Pearson sobre el coeficiente de correlación (Pearson, 1896). También afirma que esta postura fue la que dominó en los años veinte, treinta y cuarenta, tal como lo ejemplifica la siguiente afirmación de Guilford: «*una prueba es válida para cualquier cosa con la cual se correlaciona*» (1946, p. 429). La postura pragmática para el estudio de la validez, posteriormente, incorporó el análisis factorial propuesto por Spearman (1904); Sireci (2009) narra cómo Thurstone (1932), Anastasi (1938) y Guilford (1946) tomaron estos análisis como fuente para identificar la validez de una prueba.

En la década de los cincuentas, la Asociación Americana de Psicología (APA, por sus siglas en inglés), con el fin de especificar las cualidades que debería tener un test, les pidió a varios expertos de medición una definición de *validez*, quienes propusieron organizar el concepto en cuatro tipos de acuerdo con las características de las investigaciones de validez y de sus interpretaciones (Cronbach y Meehl, 1955). Aunque, esta organización no fue apoyada por todos los estudiosos de la medición en psicología, la APA la acogió en los Estándares. A continuación, se describe la propuesta de conceptualización de validez presentada en los estándares de evaluación de esta década, basada principalmente en la postura encabezada por Cronbach, la cual ha sido una de las posturas más influyentes en la conceptualización y estudio de la validez de la medición en psicología.

Cronbach y Meehl (1955) presentan cuatro tipos de estudios de validación de un instrumento de medición, estos son: a) validez predictiva, b) validez concurrente, c) validez

de contenido, y d) validez de constructo. La *validez predictiva* y la *validez concurrente* tienen una orientación de criterio, es decir, que para realizar el estudio de validación de una prueba se necesita de otra medición que sea válida para ser correlacionada con esta; la diferencia entre ambas radica en que si el criterio es aplicado después de la administración de la prueba se está estudiando la validez predictiva, pero si el puntaje de la prueba y el resultado del criterio son determinados al mismo tiempo, se habla de validez concurrente. Esta última también es estudiada cuando se quiere que una prueba sea sustituta de otra, en este caso la prueba original y sus resultados constituyen el criterio.

El criterio debe estar definido claramente. Bechtoldt (1951) afirma que la *validez de criterio* implica que se haya aceptado el conjunto de operaciones como una definición adecuada del criterio, si no se cuenta con un criterio válido, es mejor realizar una validación de constructo; sin embargo, el criterio en muchas ocasiones no es más válido que la prueba.

La *validez de contenido* se estudia cuando se quiere establecer si los ítems son una muestra representativa del universo del atributo que el investigador está evaluando. De tal manera, que se necesita la definición del contenido de la variable por medir para poder deducir cuáles son los ítems que mejor representan la variable evaluada. Este tipo de validez se enfoca en la identificación del comportamiento involucrado en el desempeño de la prueba.

Respecto a la validez de constructo, Cronbach y Meehl (1955) afirman que, «cuando se quiere que una prueba sea interpretada como una medida de un atributo que no está definida operacionalmente, se debe realizar una validación de constructo» (p. 282), es decir, que esta debe ser estudiada cuando la definición del criterio o del contenido de la prueba no ha sido aceptada o no es la adecuada según los expertos en el área y cuando se necesitan otros tipos de evidencia que den cuenta de las interpretaciones o conclusiones que pueden ser derivadas del instrumento de medición y de su validez. En una validación de constructo se busca evidencia para defender una proposición realizada a partir de una prueba; para ello, se debe contar con varios tipos de evidencia y estos deben ser evaluadas en contexto para decidir sobre la validez de la prueba. Un constructo es definido como «un atributo postulado de las personas, que se supone es reflejado en el rendimiento de una prueba. En la validación de un test el atributo del cual nosotros hacemos inferencias en la

interpretación de la prueba es el constructo. Se supone que las personas tienen o no un nivel cualitativo de un atributo o que tienen un grado cuantitativo de un atributo» (Cronbach y Meehl, 1955, p. 283).

Esta propuesta introdujo el término de validez de constructo y guió metodológicamente el estudio de la validez por décadas; sin embargo, se convirtió en una guía operacional y se dejó de lado el estudio de las interpretaciones de las puntuaciones en los estudios de validez que era fundamental en la propuesta original. Es entonces cuando el estudio de la validez se divide en tres tipos: validez de criterio, de contenido y de constructo, como si no tuvieran aspectos en común. Este panorama siguió inquietando a los expertos que estudiaban el tema de la validez y se generaron nuevas propuestas como la de Messick (1989), la cual reúne aspectos ya planteados por Cronbach y Meehl (1955) y por los demás expertos en los años cincuenta, y trata de unificar el concepto de validez. La propuesta de Messick (1989) retoma el tema de la interpretación de las puntuaciones y lo refuerza al incorporar como aspecto central de la validez, la consideración de las consecuencias del uso de las pruebas.

Concepto de validez desde la perspectiva unificada

Según Newton y Shaw (2014), en los años setentas, los ochentas hasta finales de los noventas se trabajó en torno a la visión unificada de la conceptualización de la validez, basada en el principio de ver la validación como una investigación científica sobre el significado de los puntajes de una prueba. Específicamente el trabajo de Messick buscó dar una base a la importancia que tienen las consecuencias sociales, la evaluación ética y el valor social del uso de las pruebas en los campos de la educación y psicología. Newton y Shaw (2014) sostienen que la idea de unificación entre la ciencia y la ética con una estructura para la definición de validez, propuesta por Messick, falló, aunque sí logró unificar la ciencia de la validez. Ellos señalan que la tesis principal de Messick fue profundizar en las raíces científicas de la validez y el reconocimiento de las consecuencias del uso de las pruebas desde una aproximación filosófica de la ética. Desde muy al inicio de los trabajos de Messick tuvo presente dos cuestionamientos, uno de ellos se refería a una pregunta científica enfocada a las propiedades psicométricas de las pruebas y el otro era una pregunta ética, dirigida a evaluar las consecuencias del uso de pruebas en términos de

impacto en lo social y humano de su uso, al inicio las trató separadas (Messick y Anderson, 1974); pero después ya empezó a considerarlas no separables (Messick, 1980).

Messick también realiza un análisis muy crítico de la suficiencia de la validez de contenido, de criterio, predictiva y de constructo, resaltando la importancia de la medición en la predicción, y siguiendo la idea que existe una relación en la medida predictora y el constructo de criterio, es decir, que la evidencia dada desde un estudio correlacional no puede asumirse como una justificación empírica para el uso predictivo de una prueba sino como una respaldo de la hipótesis de la relación entre la medida predictora y el constructo de criterio. Por lo tanto, es necesario contar con argumentos que sustenten la predicción y no solo evidencias empíricas, este es el proceso que se debe buscar en las prácticas de validación. Entender las relaciones de la predicción como una de tantas relaciones que conducen a entender el significado de la medición llevan a Messick a estudiar y a detallar con mayor profundidad la importancia de la red nomológica. Messick (1980) señala que la red nomológica puede ser una forma de pensar las relaciones entre los constructos predictores y los constructos de criterio, viendo la validez de criterio como un caso especial de la validez de constructo, igualmente utiliza la idea de la red nomológica para explicar el significado de las consecuencias sociales en la validación, asumiendo que estas que tendrían una relación relevante con el constructo.

Messick (1989) afirma que las inferencias de interpretativas de los puntajes es que a partir de los resultados de una prueba se infiere un nivel de atributo y que las evidencias de las consecuencias en general serían sobre los comportamientos que se esperan de la persona, y en la medida que un nivel de atributo se usa para obtener inferencias sobre conductas esperadas, entonces se tendría que la evidencia de consecuencias se vuelve evidencia para la interpretación.

Messick según Newton y Shaw (2014) postura varias ideas que son importantes para la teoría de validez: a) lo que se valida son los resultados de un procedimiento de medición y no una prueba, b) cuando se habla de validez se asume que ya se está midiendo, es decir, que la medida antecede a la validez, c) toda la validez es de constructo, d) el concepto de validez tiene diferentes facetas, e) el procedimiento de validación es una investigación científica, f) el rasgo principal de la validez de constructo es poder descartar posibles

hipótesis alternativas, g) la validez es una cuestión de grado, no es de todo o nada y finalmente, e) la validación es un proceso de mejoramiento continuo.

De tal manera que Messick (1989) define la validez como «*un juicio integrado del grado en el cual la evidencia empírica y la argumentación teórica soportan que las inferencias y las acciones basadas en los puntajes de una prueba o de otro modo de evaluación son adecuados y apropiados*» (p.13). Al igual que postula la matriz progresiva (Figura 1) como una representación para estudiar las facetas de la validez, esta representación estructura la forma de realizar el proceso de validación.

	Interpretación del test	**Uso de test**
Bases de evidencia	Validez de constructo (VC)	VC + Relevancia y utilidad (R/U)
Bases de consecuencias	VC + Implicaciones de valor (IV)	VC + IV + R/U Consecuencias sociales

Figura 1. Matriz Progresiva

La estructura de validación inicia con la celda de *validez de constructo* se ubican cinco aspectos de la validez: los aspectos de contenido, referido a la relevancia y a representatividad del dominio; los aspectos sustantivos, que se refiere a las teorías que sustentan la prueba y a los modelos de procesos de respuesta; los aspectos estructurales, referidos a los modelos de calificación, en la medida que reflejen la estructura del dominio y de las tareas; los aspectos de generalizabilidad, referidos a la generalización entre momentos de aplicación, tareas, observadores, etc.; y finalmente, los aspectos externos, que tienen que ver con las correlaciones con otras variables (convergente o discriminante). Después se pueden seguir dos vías: una que es la que lleva a la celda de *implicaciones de valor* tiene que ver con los juicios valorativos relacionados con el significado del puntaje y la otra que conduce a la celda de la *relevancia y utilidad*, la cual abarca la validez de constructo y la recolección de evidencia específica para un propósito particular, y siendo el caso también puede entrar la validez de criterio. Y finalmente la celda de *consecuencias del uso*, apuntan al valor funcional apreciación de las consecuencias sociales que tiene el uso

de la prueba, enfatizando que las consecuencias adversas de la prueba no la invalidan, al menos que estas afecten la interpretación de los puntajes.

Las propuestas de Messick respecto a la validez y a la validación influyeron en los trabajos académicos posteriores en la teoría de validez y en las recomendaciones técnicas para el diseño y uso de las pruebas. Su influencia se nota en la publicación de los Estándares para el uso de pruebas en la educación y la psicología desde 1985 hasta la actualidad. Aunque las recomendaciones técnicas y Estándares previos habían sido influenciados por la idea de la información de validez indica el grado en que una prueba es capaz de cumplir con ciertos requisitos, estos requisitos eran los cuatro tipos de validez (APA, 1955); y después de la discusión entre varios profesionales, académicos y usuarios involucrados en la evaluación por medio de pruebas se empezó a hablar de aspectos y categorías de la validez, también la definición de *validez,* Sireci (2009, 2013) describe cómo desde los Estándares de 1974 (APA, AERA, NCME, 1974) comenzó a hablarse de la validez referida a las interpretaciones de los puntajes de las pruebas y no a la prueba en sí misma. También señala como en los Estándares de 1985 la *validez* se refiere a lo apropiado, significativo y útil de una inferencia específica derivada de un puntaje de un test, también que este es un concepto unitario y que, aunque la evidencia se puede recolectar de muchas maneras, la validez siempre se refiere al grado en el cual la evidencia respalda las inferencias hechas a partir de los puntajes para un uso específico, más no para la prueba en sí misma.

En los Estándares de 1999 y de 2014 se define la validez como el grado en que la evidencia y la teoría respaldan las interpretaciones de los puntajes de la prueba para los usos propuestos. En estas dos versiones que varían en aspectos teóricos y metodológicos respecto a sus versiones antecesoras, siguen considerando el concepto de la validez unificado y proponen una estructura de validación referida a cinco fuentes de evidencia de validez, y la acumulación de estas evidencia respaldan la interpretación de los puntajes para un uso propuesto, estas cinco fuentes de evidencia son evidencias basadas en a) el contenido de la prueba, referido b) los procesos de respuesta, c) estructura interna, d) relaciones con otras variables y e) consecuencias del uso de la prueba.

Sireci (2009, 2013) explica que, en los Estándares de 1999, la evidencia basada en contenido se entiende como la evidencia tradicional de validez de contenido, en la que se revisa los ítems de la prueba y las especificaciones de la prueba y también se enfoca al estudio de la coherencia entre las estructuras de currículo, uso de la prueba e instrucciones, particularmente en el ambiente educativo. Respecto a la evidencia basada en los procesos de respuesta se refieren a la evidencia del ajuste entre el constructo y las respuestas de los examinados, estas evidencias pueden ser recolectadas por medio de entrevistas a los evaluados respecto a sus respuestas, observaciones sistemáticas de conductas asociadas a responder una prueba, evaluación de criterio usada para evaluar pruebas de desempeño, análisis del tiempo de respuesta de los ítems y análisis de procesos de razonamiento usado para contestar pruebas. En cuanto a la evidencia basada en la estructura interna, es la evidencia que responde al análisis estadístico de la prueba y de los ítems para investigar sobre las dimensiones medidas en una evaluación, algunos procedimientos para recolectar dichas evidencias son análisis factoriales tanto exploratorios como confirmatorios, análisis de ítems (TCT-IRT) y escalamiento multidimensional. Respecto a la evidencia basada en las relaciones con otras variables se refiere a la evidencia a las formas tradicionales de analizar la validez relacionada con el criterio, tanto concurrente como predictiva en los estudios de validez y estudio multirasgo-multimétodo. Y finalmente, la evidencia basada en las consecuencias del uso de la prueba se refiere a la evaluación de las consecuencias previstas y no previstas asociadas con el programa de evaluación, ejemplos de estas son las evaluaciones que incluyen el estudio de impacto adverso, efectos de la instrucción en la evaluación, efectos de los usos de las pruebas en la deserción escolar. Elosua (2003) y Sireci (2013) describen en datalle los procedmientos metodológicos involucrados en la recolección de las evidencias internas y externas para cada una de las fuentes de validez.

Los Estándares de 2014 realizan algunos cambios respecto a la versión de 1999, sin embargo la que más nos concierne se refiere a la quinta fuente de evidencia de validez, referida ahora a la evidencia para la validez y consecuencias del uso de las pruebas, en esta se especifican más detalles sobre las interpretaciones y usos de los puntajes propuestos por parte de los constructores de pruebas, las afirmaciones acerca el uso de la prueba que no necesariamente están basadas en la interpretación de los puntajes y las consecuencias que

no son previstas por el uso de la prueba, además de anexar un capítulo sobre el análisis de consecuencias en el campo educativo al final de los Estándares.

Si bien el trabajo de compilación e integración realizado por las diferentes asociaciones involucradas en la publicación de los Estándares, que reúne a académicos, profesionales, usuarios, comercializadores, políticos, entre otras por lograr una definición unificada de validez y de validación, se sigue presentando desacuerdos relacionados con la definición de la validez, estos puntos han sido foco de atención de diversos académicos a nivel mundial en la última década y media.

Discusión del concepto de validez de final del siglo XX e inicios del siglo XXI

En esta sección se abordará la discusión sobre el concepto de validez que se ha llevado a cabo a finales de los años noventa y los años que han transcurrido de los dos mil. Como guía para la presentación de las posturas de los académicos se tendrá en cuenta el trabajo realizado por Newton y Shaw (2016) en el que buscan persuadir a el lector de que es interesante e importante conocer por qué no se ha alcanzado un consenso en la definición de validez y también considerar si es plausible lograr este consenso. Respecto al consenso, los autores señalan que se podría decir que la mejor forma de usar la palabra *validez* ya se alcanzó, o que, ya estamos más cerca de alcanzar un consenso que antes o que a pesar de las diferencias de terminología, las definiciones que actualmente ya existen dicen más o menos lo mismo. Sin embargo, son enfáticos en demostrar que ninguna de estas afirmaciones es correcta y que la situación se describe mejor como un enfrentamiento entre académicos que abogan por usos radicalmente diferentes. También afirman que el debate se ha venido incrementando en los últimos años y la discusión tiene muchos puntos a tratar, aunque los dos más importantes para responder son: ¿Qué debe abarcar la validez? Y ¿a qué se le debe aplicar la validez? Estos autores clasificaron las posturas de los académicos frente a estas preguntas en tradicionalistas, liberales, conservadoras y ultraconservadora, para ilustrar la naturaleza y características del desacuerdo respecto al concepto de *validez*.

Tradicionalistas. En esta categoría se encuentran los académicos que consideran que la *validez* debe abarcar la medida, entendida como la interpretación de los puntajes, y la predicción, referida al uso de los puntajes. Es decir, siguen la propuesta de la primera

edición de los Estándares, los trabajos de Cronbach (1971) y de Messick (1989), en los que se resalta que es inviable pensar la interpretación de los puntajes y el uso de los puntajes como aspectos diferentes. Están de acuerdo con que la evaluación por medio de pruebas no es ciencia pura, que la validez es inherentemente pragmática y que las pruebas no son creadas en el vacío, sino son creadas con un propósito y los puntajes no son generados simplemente para ser interpretados sino generalmente para ser usados. Newton y Shaw (2016) señalan particularmente que, desde la perspectiva tradicionalista, casi se podría afirmar que es el uso lo que se debe validar, y por esto para ellos es importante hablar de la utilidad de las pruebas, la cual va más allá de la mirada técnica, porque destaca los intereses económicos, sociales, y éticos, referidos a términos de costo beneficio y que este componente sugiere que los tradicionalistas deben ser capaces de distinguir qué tipo de juicios de valor requieren para establecer la utilidad de la prueba y prestar mayor atención a las consideraciones éticas y sociales respecto a las de la utilidad económica y las de justicia moral, que ya han sido foco de atención.

Uno de los representantes de esta postura es Stephen Sireci (2007, 2009, 2013, 2016a), quien en sus trabajos ha señalado que la validez no es una propiedad de la prueba, sino que se refiere al uso que se haga de esta para un propósito particular, que para evaluar la utilidad y qué tan adecuada es una prueba se necesitan muchas fuentes de evidencia y que si el uso de una prueba debe ser sustentable para un propósito particular, se deben presentar pruebas suficientes para defender el uso de la prueba para dicho fin. Está de acuerdo con que no se puede evaluar un test independiente de su propósito y que se necesita saber el fin de los puntajes del test para realizar la validación de la prueba.

Sireci (2016), a través de un recuento histórico, ilustra cómo en la definición de validez dada en Estándares siempre se han visto dos nociones importantes a pesar de que en las primeras ediciones se presentara una definición fragmentada de la validez; la primera, que la validez no es una propiedad inherente a las pruebas y la segunda, que la validación debe enfocarse en los propósitos del test. También señala que la cuarta edición de los Estándares (AERA et ál., 1985) provee un ejemplo claro de una definición confusa, ya que si bien hace más énfasis en las inferencias que en los propósitos y usos de la prueba, también afirma que las inferencias sobre los usos también deben ser validados y no el test como tal. Sireci

enfatiza que esta definición poco clara condujo a los desacuerdos entre los académicos y usuarios que afirman que la validez se refiere a las interpretaciones (p. ej. Cizek, 2012), con aquellos que afirman que es de los usos (p. ej. Sireci, 2013) y aquellos que dicen que la validez se refiere a ambas (p. ej. Kane, 2013).

Para Sireci, la idea de que la validez se refiere solo a la interpretación de los puntajes ganó adeptos en la última parte del siglo XX. Esto lo muestra mediante un recorrido histórico desde la definición de validez dada por Messick (1989, p. 13) «*la validez es un juicio valorativo integrado del grado en el cual la evidencia empírica y los argumentos teóricos respaldan que las inferencias y las acciones basadas en los puntajes de las pruebas y otras formas de evaluación sean adecuadas y apropiadas*», e ilustra cómo esta definición unificada genera confusión sobre el concepto al involucrar en la validez dos lados. Un lado es la fuente de justificación de la evaluación, basándose en la valoración de las evidencias o consecuencias. El otro lado es la función o resultado de la evaluación, siendo la interpretación o el uso. Así es como la interpretación de esta distinción creó confusiones entre los académicos; sin embargo, para él es claro que no puede verse la validez solo teniendo en cuenta la interpretación de los puntajes, sino que también deben estudiarse los usos de los mismos.

Sireci (2007, 2009, 2016a, 2016b) está de acuerdo en que es necesario conocer si lo que la prueba mide es lo que se propone medir (Borsboom, 2016), con la importancia de la validación de contenido de los test (Lissitz y Samuelsen, 2007), con que la interpretación de los puntajes es parte de la validación y parte de la validez; pero también es reiterativo en afirmar que esto no es suficiente para defender el uso de la prueba para un propósito particular.

Sin embargo, Sireci (2016a) señala que la falta de claridad en la definición de la validez dada en la cuarta edición de los Estándares (AERA et ál., 1985) fue resuelta en la edición actual de los Estándares (AERA et ál., 2014), ya que en estos últimos queda explícito que los usos de los test son inseparables de las interpretaciones. El autor sostiene que en la realidad no se puede tener una interpretación de los puntajes del test sin conocer el uso que se le dará a los puntajes. Es decir, que no es necesario separar la definición de la validez dada por los Estándares en diferentes componentes, que una definición suficiente de la

validez incluye tanto las interpretaciones de los puntajes como los usos de estos puntajes. Un test puede ser validado teóricamente, pero es importante saber para qué es validado, ya que se necesitan validar las acciones y los usos, no solo interpretaciones que no están situadas en ningún lugar.

Por lo tanto, Sireci (2016a, 2106b) propone acoger la definición propuesta por los Estándares (AERA et ál., 2014), ya que esta definición integra las interpretaciones y los usos, además de dar buenos fundamentos para guiar el proceso de validación. También que se deje atrás la discusión si los usos están o no dentro de la validez, y que mejor se dediquen esfuerzos a buscar evidencias para el uso de las pruebas. Está de acuerdo con que las consecuencias de los usos de los test sean parte de la validez y su trabajo es y será enfocado principalmente a buscar evidencias para respaldar el uso del test para un propósito específico, por medio de varias fuentes de evidencia, unas basadas en la teoría o juicios subjetivos y otras basadas en análisis estadísticos.

Liberales. Newton y Shaw (2016) presentan la perspectiva liberal como aquella en la que el centro de atención es la discusión sobre la importancia de las consecuencias para la validez y la validación. Para los académicos clasificados en esta postura, no tener en cuenta las consecuencias no solo es insuficiente sino irresponsable. La postura liberal se puede ver como una extensión de la postura tradicional ya que mientras que los tradicionalistas ampliaron la definición clásica de validez para incorporar la medición (la interpretación de los puntajes) y la predicción (el uso de los puntajes), los liberales la amplían para abarcar además las consecuencias intencionadas y las no intencionadas del uso de los puntajes de los test. Esta postura se puede observar en los primeros trabajos de Messick (1980) al igual que en Cronbach (1988); en la actualidad se puede identificar en los trabajos de Moss (2016, 1998), o Shepard (1997, 2016); sin embargo, no es sencillo identificar representantes de esta postura por los múltiples matices presentes en sus ideas.

Una de las académicas que han expresado su postura liberal «moderada», según ella misma, es Pamela Moss (1998, 2006, 2007, 2016). En sus trabajos sobre la validez de la evaluación en el contexto educativo, ella ha mostrado cómo la discusión sobre las consecuencias ha estado presente en la definición de la validez desde sus inicios, por ejemplo, en los trabajos de Cronbach (1980, 1988), Messick (1989) y de Shepard (1993,

1997). Para Moss, cada uno de estos teóricos articula una perspectiva diferente sobre la relación entre validez y consecuencias, y construye un argumento sobre bases algo diferentes; estas perspectivas pueden conducir a conclusiones diferentes sobre el grado de validez asociado con una interpretación o uso dado de una prueba. Ella sostiene que la cuestión de incorporar la consideración de las consecuencias en la definición de validez no es sólo filosófica, sino que tiene repercusiones éticas, políticas y económicas reales.

Moss (1998) argumenta que es importante tener en cuenta las consecuencias de la evaluación en la definición de validez en la naturaleza reflexiva del conocimiento social formulada por científicos sociales como Thompson (1990). Ella sostiene que, en la medida en que las prácticas de evaluación realizadas por los académicos cambian la realidad social que estudian, el estudio de las consecuencias se convierte en un aspecto esencial de la validez. Este sería el caso incluso para quienes optan por limitar el alcance de la validez a una interpretación basada en la prueba.

Uno de los argumentos de Moss (1998, 2016) es que la interpretación fija de los puntajes de una prueba que pueden tener los académicos que investigan la validez no se mantiene constante entre los contextos locales de evaluación. Estas interpretaciones dependen de los recursos disponibles y de las circunstancias sociohistóricas de los individuos involucrados en el sistema de evaluación; por lo tanto, las consecuencias de la difusión de estos mensajes dependen de cómo los individuos incorporan estos mensajes en su vida cotidiana y de cómo estos mensajes afectan la forma en que ellos mismos se ven a sí mismos y a los demás. Ella insiste en que es importante darle valor a la interpretación de los resultados, no solo en términos del significado pretendido del mensaje, sino también en términos de sus implicaciones, que no siempre son intencionadas, en los lectores y en la naturaleza de su conocimiento. Por lo cual lo que se necesitaría es estudiar el discurso real y las acciones que ocurren alrededor de las pruebas, informes, programas y demás prácticas derivadas de la evaluación por medio de pruebas.

Moss (1998) sostiene que, la idea de que las consecuencias son parte esencial de la *validez* puede salpicar la búsqueda de consenso en la teoría de la validez en relación a las interpretaciones y usos de los puntajes de la prueba, pero que su llamado es a que en el consenso en la teoría de la *validez* también se tenga en cuenta las prácticas generales de

evaluación, su relación dialéctica entre los productos y prácticas de evaluación, y la realidad social, que es representada y transformada continuamente.

Moss et ál. (2006, 2016) afirma que la teoría de la *validez* necesita sustentar aproximaciones conceptuales que ayuden a los educadores a conectar los datos generados por las pruebas con sus prácticas, a generar explicaciones y a explorar soluciones. Ella propone que la teoría de la *validez* cambie el foco de atención, centrado en las interpretaciones intencionadas y los usos de las pruebas que orientas a los constructores y diseñadores de pruebas, por un enfoque hacia las interpretaciones reales, las decisiones y las acciones que sirven a propósitos específicos de los usuarios en contextos particulares, que involucre a los investigadores y a los usuarios de las investigaciones de otros. Este trabajo interdisciplinario robustecería la teoría de la validez sobre un uso conceptual de los puntajes de las pruebas al orientar las preguntas específicas de una organización para que esta tenga la capacidad de hacer un uso adecuado de los datos y al darle valor a las pruebas para mejorar la práctica educativa.

Otra académica que ha argumentado sobre la importancia de tener en cuenta las consecuencias en la definición de *validez* es Lorrie Shepard (1993, 1997, 2016). Shepard (1997) también señala que la importancia de las consecuencias del uso de las pruebas como parte de la validez no fue algo una novedad introducida por Messick. De acuerdo con su revisión, desde Curenton (1951) se afirma que la validez no es de la prueba, sino que debe ser evaluada de nuevo en cada aplicación, ya que el significado de lo que se está midiendo por la misma prueba puede variar dependiendo de las experiencias del grupo evaluado. Shepard también resalta cómo Cronbach (1971) incluyó dentro del ámbito de validez los estudios de evaluación de decisiones y acciones basadas en las puntuaciones de los exámenes, así como las interpretaciones descriptivas; una vez que se incluye la solidez de las decisiones basadas en pruebas como parte de la validez -no sólo las descripciones o interpretaciones sin acciones-, se está obligado a pensar en los efectos o consecuencias. Shepard (1997) añade que la matriz de Messick (1989) es un error, pues introduce la posibilidad de evaluar la validez de constructo de una interpretación de prueba sin considerar el uso de la prueba mientras no se pretenda su uso.

Para Shepard (1997), en una investigación de validez, no sólo se expresa una preferencia personal por las consecuencias deseables o no, sino que las consecuencias se evalúan en términos del significado de construcción deseado. Así que los efectos, intencionados o no, se reflejan de nuevo en el significado de la puntuación de la prueba y hacen parte de la validez. Para cada uso de una prueba, debe haber un marco de validez o una red conceptual que incluye los efectos previstos y los efectos secundarios posibles o eventualmente identificados. Cuando se adopta una prueba existente para un nuevo propósito, se requiere una nueva evaluación de validez, aun cuando algunos datos existentes puedan ser relevantes. En este sentido, ella concluye que se deben incluir las consecuencias en los estudios de validez. Shepard (1993) y Kane (1992, 2016) han sugerido algunas maneras de utilizar un enfoque basado en argumentos para priorizar las preguntas sobre la *validez*.

Respecto al consenso, Shepard (2016) opina que el desacuerdo sobre la definición de *validez* no es necesariamente un problema en el campo científico, siempre y cuando se tenga clara la naturaleza del desacuerdo y se observe cómo las diferencias en las concepciones conducen a su vez a diferencias en métodos y hallazgos. También sostiene que el consenso en la definición alcanzado en los Estándares (2014) es el mejor logrado hasta el momento, ya que involucra las posturas de un comité de expertos, la retroalimentación de la revisión y la aprobación de muchas organizaciones que trabajan en el área. Para ella, tener una definición oficial da claridad incluso para quienes están en desacuerdo con la misma porque sirve de criterio y los obliga a conocer cómo y por qué ellos difieren de este criterio. De igual forma, resalta la importancia que tiene esta definición de *validez* al considerar las interpretaciones y el uso como un todo y no como una secuencia de pasos; cuando los Estándares (2014) se refieren a «la evidencia y la teoría», están pidiendo no solo que el test esté bien diseñado según la teoría, sino que la evidencia que sea recolectada verifique que sí está funcionando como se quiere que funcione.

Conservadores. Según Newton y Shaw (2016) esta perspectiva representa a quienes consideran que la validez es el grado en que la prueba mide lo que dice medir (Ruch, 1924. p. 13). Para los proponentes de esta perspectiva, como Cizek (2012), la *validez* es un

concepto científico no un concepto pragmático; por consiguiente, la validez es un concepto que solo puede ser aplicado a la medida y a los procedimientos de medición, y es un error categorial validar predicciones, usos de los resultados y el proceso de toma de decisiones. Por lo tanto, la interpretación de los puntajes y el uso de los test son aspectos incompatibles. También consideran que la *validez* se debe centrar en la calidad de la medida, y esta abarca varios conceptos relacionados pero que se distinguen entre sí tales como la confiabilidad/precisión, la dimensionalidad y el sesgo.

Cizek, Rosenberg y Koons (2008) mostraron cómo una de las fuentes de validez, la validez de las consecuencias del uso de la prueba, la cual era rutinariamente ignorada en las presentaciones de la evidencia de validez publicadas en el *Mental Measurement Yearbook*. En el estudio de Cizek, Bowen y Church (2010a) se replica el estudio de Cizek et ál. (2008) utilizando otras fuentes de datos, una muestra de revistas de políticas de medición y pruebas aplicadas en los últimos 10 años que incluye más de 2.400 artículos, de los cuales más de mil trataron sobre la validez. En este estudio no se encontraron casos en los cuales se presentara o se describiera evidencia de validez basada en las consecuencias de la prueba. En este estudio también se hizo una revisión de los programas de reuniones anuales de las tres asociaciones profesionales patrocinadoras de los Estándares, para los años 2007 y 2008, en este tampoco se encontró ninguna investigación en la que se recogiera o resumiera información sobre las consecuencias de la prueba como parte de una investigación de validez. Estos resultados confirman los hallazgos reportados anteriormente por Cizek et ál. (2008). Por lo tanto, los autores concluyen que las consecuencias de las pruebas son rutinariamente ignoradas como una fuente de evidencia de validez y que la teoría de la validez tiene un fallo al incluir la validez consecuencial en su definición.

Cizek (2010b, 2012, 2016a, 2016b), respecto a la falla en el concepto de validez, afirma que esta radica en que la validación de las interpretaciones y la justificación de un uso específico de la prueba son dos aspectos diferentes, separados y que NO se pueden combinar. Es decir, que el "juicio integrado evaluativo" propuesto por Messick (1989) nunca se ha visto. Según Cizek (2010a, 2010b, 2012, 2016a) y Cizek et ál. (2010a) no es posible hacer una síntesis de la interpretación y de las consecuencias en una sola conclusión. Lo que si es viable es obtener conclusiones sobre cada una, pero no de forma

integral; por lo tanto, Cizek (2010a, 2012, 2016a) afirma que la validez de las consecuencias no existe, pero las consecuencias del uso de las pruebas sí, y que estas son tan importantes como las interpretaciones de los puntajes de la prueba. Por lo que se necesita una separación de la recopilación de evidencias que influyen en el significado de las puntuaciones de los test de la recolección de evidencias que llevan a la conveniencia de utilizar la prueba. Para Cizek, separar la validación de las inferencias de la puntuación de la justificación del uso de la prueba ayudan a aumentar el rigor metodológico para ambas, y se robustece y aclara el concepto de validez. Esta conclusión se basa en los resultados contundentes de los estudios de Cizek (2008, 2010a, 2010b), en los que en ningún caso la evidencia basada en las consecuencias de la prueba se ha llevado a estudiado, ni por sí misma ni con otras fuentes de evidencia de validez para dar un ejemplo de un caso coherente y general a favor o en contra de la validez de las puntuaciones derivadas de un instrumento en el contexto de educación o de psicología.

Cizek (2012, 2016a, 2016b) propone la validación de las interpretaciones y la justificación del uso del test como dos procesos separados. Sin embargo, los procesos de validación y de la justificación pueden verse como una secuencia lógica; primero, con la validación, se obtiene la evidencia para apoyar la interpretación que se pretende del puntaje de la prueba, esta se va recolectando durante el desarrollo de la prueba y en la evaluación, luego sigue la obtención de la evidencia que justifique el uso de la prueba, para esto el significado de las puntuaciones deben estar sustentadas aunque la justificación puede empezar cuando la prueba ha sido aplicada e incluye información sobre las consecuencias del uso de la prueba. Todo esto alimenta los juicios de valor respecto a la prueba. Los procesos de validación y de justificación constantemente interactúan, ya que la evidencia de la validez de una inferencia es una condición necesaria pero insuficiente para recomendar una justificación del uso de una prueba.

Respecto a los Estándares (2014), Cizek (2016a) afirma que cuatro de las cinco fuentes de validez propuestas son apropiadas; sin embargo, la quinta: «*Evidencia de la validez de las consecuencias del uso del test*» no pertenece a este grupo. También señala la importancia de desarrollar fuentes de evidencia robustas para justificar el uso de las pruebas, tales como las que se tienen para sustentar las interpretaciones de los puntajes de

las pruebas. Adicionalmente, resalta los trabajos relacionados con la justificación del uso de las pruebas en el campo de '*programme evaluation'*, desarrollando estrategias sistemáticas de investigación como la evaluación de necesidades, aproximaciones de costo-beneficio, análisis de costo-utilidad e investigaciones de costo–efectividad, investigaciones en las que tienen en cuenta a todos los involucrados en el sistema de evaluación.

Ultra conservadores. Newton y Shaw (2016) enmarcan en esta perspectiva a aquellos autores que recomiendan que se vuelva al concepto clásico de validez. Borsboom et ál. (2004, 2009, 2012, 2013 2016) argumentan que la validez se debe referir a lo que toda medida debe ser, es decir, que lo que se está tratando de medir cause variaciones en los resultados de los procedimientos de medida. Si esta relación causal existe, entonces, el procedimiento es válido para medir lo que dice medir, y si no se encuentra esta relación causal entonces no es válido. Para este grupo, a diferencia de los conservadores, la validez es independiente de otros conceptos tales como la confiabilidad/precisión, la dimensionalidad y el sesgo, ya que la validez no sería un problema metodológico sino sustancial-ontológico, y no son vistos como fuentes de evidencia de validez. Es decir, un instrumento podría ser válido y no confiable, o ser válido y a la vez presentar sesgo.

Borsboom et ál. (2004, 2009, 2016) sustentan la idea de que es necesario validar el test y no las interpretaciones. Para ellos, validar las interpretaciones es una idea equivocada que se deriva de la validez de constructo, ya que la validez es una propiedad de los test, no de las interpretaciones, ni de los puntajes y tampoco de los constructos.

Borsboom et ál. (2004, 2009) afirman que la teoría de la validez ha de basarse más en un componente ontológico, es decir en cómo son las cosas, más que en uno epistemológico, que responde a cómo conocemos las cosas; de tal manera que la validez se debe centrar en la referencia y la causalidad y no en el significado ni en la correlación. En esta medida, tanto los aspectos epistemológicos como los relacionados con las consecuencias del uso de una prueba no son relevantes cuando se define la validez, ya que lo que le debe interesar a la validez es solamente si el instrumento tiene la capacidad de capturar la variación en un atributo especifico, es decir, si el instrumento mide lo que dice medir.

Por otra parte, respecto a la definición de grado en el concepto de validez, Borsboom et ál. (2004, 2009) argumentan que esta no aplica en su postura, pues para el estudio de la

validez se debe evaluar si se presentan dos condiciones; la primera, es si el atributo existe y la segunda, es si la variación del atributo causa la variación en las puntuaciones de la prueba. El resultado de esta evaluación solo se puede responder categóricamente, es decir si sí se presentan o no, tal como se hace con el concepto de verdad y esta se puede resumir en si el test mide lo que dice medir.

Borsboom et ál (2004, 2009, 2012, 2016) consideran que, si bien el interés de la validez debe centrarse en el carácter ontólogico de la medición, los aspectos epistemológicos y los éticos también son importantes para la validación de instrumentos. Ya que los intereses científicos y políticos tienen diferentes objetivos, es decir, mientras los científicos se interesan por los valores de verdad de las afirmaciones, los políticos se centran en el juicio moral de estas, entonces respecto a las pruebas, responder las preguntas sobre la justificación de las pruebas (político) no resuelve ni la pregunta sobre la validez de la prueba (científico) ni vice versa, simplemente porque son dos cosas diferentes. Estos dos puntos se deben trabajar por separado y evitar combinarlos pues crea mucha confusión.

Por otra parte, Borsboom (2012) afirma que la posición de no requerir que las interpretaciones de la medición sean correctas, sino solamente que el argumento con el que se las justifica es suficientemente fuerte para sostenerlas, es una posición legalista. Bajo esta posición se defiende el uso de las pruebas ante usuarios, jueces, políticos, lo cual es funcional para la sociedad; pero esta le falla a los intereses de los científicos investigadores. Esta visión es mucho más pragmática que la estudiada desde la psicología científica, ya que facilita y agiliza los procesos de certificación de las instituciones que regulan el uso de los test. Por lo tanto, esta visión de validez no concuerda con la postura sobre la validez desarrollada por Bosrboom, en la que prima el valor de verdad respecto a si el test mide o no lo que dice medir.

El desarrollo de la teoría de la validez necesita psicometría, filosofía de la ciencia y fundamentos de la teoría psicológica, necesita investigación y el trabajo en conjunto e integración de teoría psicológica, conocimentos en construcción de pruebas y de análisis de datos para resolver problema de la validez con gran detalle (Borsboom, 2009). Se necesita del concepto de validez para comunicarnos entre nosotros respecto a si una prueba mide lo

que se propone o no, y también por que el desarrollo de la investigación de la teoría de la validez estimula la crítica y el desarrollo científico (Borsboom, 2012).

También es importante anotar que la definición dada por esta postura, en la que la noción de validez se entiende como "*que el test mida lo que se pretende medir"* (Buckingham et al., 1921; Ruch, 1924; Garret, 1937), se sigue presentando como la más común y trasmitida por libros de textos básicos de psicometría (p. ej. Allen & Yen, 1979/2002, p. 95; Brown, 1980, p. 40; Martínez, 1996, p. 37; Muñiz; 1996, p. 55). Es sobre esta que normalmente se apoyan las interpretaciones de los puntajes de las pruebas (Muñiz, 1996; Sireci, 2009), y si bien, en esta definición es central la acción de medir; se observa en estos mismos textos, que todo lo que implica la medición es normalmente evitado (Martínez, 1996), o se ha limita a reiterar que medir es únicamente "*asignar números a objetos o eventos de acuerdo con reglas"* (Stevens, 1946, p. 677; Brown, 1980); esto también involucra el proceso denominado *escalamiento*, que es por el cual se definen las reglas de asignación de números, y se subsumirá la obtención de la medida del atributo psicológico "*que tiene lugar cuando se asigna un valor cuantitativo a la muestra de conductas recogidas"* (Martínez, p. 29).

Markus y Borsboom (2013) identifican cuatro posturas filosóficas sobre lo que puede entenderse como medición, las cuales resuelven el problema planteado para el escalamiento. La primera postura tiene una orientación realista y la denominan "Teoría clásica de la medición" (distinta a Teoría Clásica de los Test), afín a la perspectiva presentada por Bunge (1973) y defendida en psicología por Michell (1999); la postura operacionalista derivada de Stevens (1946); la postura representacional presentada por Suppes y Zinnes (1963); y la perspectiva de variables latentes defendida por estos autores (Markus & Borsboom, 2013).

En la *teoría clásica de la medición*, medir se define como "*atribuir valores concretos a variables(s) numéricas(s) de un concepto cuantitativo sobre la base de la observación* (Bunge, 1985, p. 769)". En esta definición, además del problema de asignar valores se requiere que los mismos correspondan a variables numéricas y que sean asignadas a conceptos cuantitativos; estos dos requerimientos son adicionales a la existencia de reglas claras y sistemáticas para llevar a cabo la asignación.

De acuerdo con Bunge (1973, 1985), es primordial que se lleve a cabo la definición del concepto cuantitativo como requisito y como antecedente de cualquier intento de medición. Esto no quiere decir que no exista un cuerpo de conocimiento empírico que haya establecido relaciones cualitativas sobre el cual se logre posteriormente definir el concepto cuantitativo. Esto enlaza directamente la definición dada por Bunge con la que toma Michell (1999) como definición clásica de medición: «*descubrimiento o estimación de la razón de alguna magnitud de un atributo cuantitativo y una unidad*» (Michell, 1999, p. 14). De este modo, la definición clásica de medición implica, por una parte, una proposición sobre un atributo de un objeto tal que tiene se puede definir una función con ciertos valores numéricos y una operación empírica de observación por la cual se atribuye o estima ese valor para un objeto concreto. Sobre el primer aspecto, Bunge (1973, 1985) hace énfasis en que la definición del concepto cuantitativo es lógicamente anterior a cualquier medición, mientras que Michell (1999, 2009) hace énfasis en que la proposición (que como tal puede ser verdadera o falsa) representa una hipótesis sobre la naturaleza cuantitativa del atributo, que es necesario y posible verificar empíricamente para poder dar por cumplido el requerimiento lógico de cuantificación.

Por último, en la definición aportada por Michell (1999), se hace explícito adicionalmente que la medición, de acuerdo con la definición clásica, requiere del establecimiento de una escala para la magnitud medida, y que para ello es necesario definir una unidad de medida.

La aproximación basada en argumentos. Kane (2016) define la validez como «qué tanto las interpretaciones y los usos de los puntajes están justificados. Esta justificación requiere un análisis conceptual de la coherencia y la completitud de las aseveraciones y análisis empíricos de las inferencias y supuestos inherentes a estas aseveraciones» (Kane, 2016. p.198). Aunque su propuesta contribuye a la construcción de la teoría de la validez su aporte no consiste en fundamentar el concepto de validez sino en proponer una estructura que guíe los procesos de validación (Newton y Shaw, 2014). Para esto Kane (1992, 2001, 2013, 2016) ha planteado *la aproximación basada en argumentos* (IUA's por sus siglas en inglés), la cual consiste en la formulación de una estructura para realizar un proceso adecuado de validación atendiendo a tres condiciones importantes: (a) que la validación

necesita un programa de validación mucho más completo que un solo estudio de validez, (b) que en el proceso de validación se necesita especificar detalladamente las interpretaciones propuestas; y (c) que es vital comprobar los supuestos que respaldan cada una de las interpretaciones propuestas, tanto a nivel lógico como empírico, así como probar las hipótesis contra hipótesis igualmente viables.

Según Kane (2006, 2013), para validar una interpretación deseada se necesita tener claridad de lo que se va a inferir, y una forma para definir con claridad estas interpretaciones e inferencias es por medio de la *aproximación basada en argumentos*; ya que esta permite que las interpretaciones y los usos propuestos de los puntajes de las pruebas sean descritos con mayor detalle en forma de un argumento, desde los desempeños observados en la prueba hasta las conclusiones basada en estas. Esta estructura tiene en cuenta dos pasos: el primer paso es que se debe especificar la red de inferencias y supuestos que llevan desde el desempeño en la prueba hasta las conclusiones y decisiones basadas en estos desempeños, es decir, especificar las interpretaciones y los usos de los puntajes del test en forma de argumento o IUA; y el segundo paso es evaluar de manera crítica el argumento o IUA, este es conocido como *validez argumentativa*, la cual consiste en evaluar la coherencia, la completitud y la plausibilidad de las aseveraciones planteadas.

Con la *aproximación basada en argumentos* principalmente se busca que las interpretaciones y usos propuestos tengan sentido y sean sustentados con la evidencia apropiada. Esta estructura si bien permite utilizar una amplia gama de interpretaciones y usos también exige que cada aseveración que se vaya a hacer tenga un respaldo teórico y empírico robusto (Kane 2001, 2013, 2016).

En relación a las consecuencias en el concepto de validez, Kane (2016) plantea que la pregunta no es si las consecuencias juegan un papel en la validez sino cuál es la naturaleza y el objetivo del rol de las consecuencias y cuáles tipos de consecuencias se deben considerar. Para responder esta inquietud, él plantea tres posibilidades para explicar el papel de las consecuencias no deseadas en validez: un modelo de solo la interpretación, un modelo de las consecuencias como indicadores y un modelo de interpretación y usos.

La posibilidad de un modelo de solo la interpretación consiste en realizar la validación solo de las interpretaciones de los puntajes dejando de lado la evaluación de las

consecuencias. Respecto a esta opción Kane (2016) nombra tres desventajas: la primera es que esta opción es muy simple y permite que con una interpretación válida se tomen decisiones muy pobres y que una consecuencia negativa o una política mal planteada no invalide dicha interpretación. La segunda es que en evaluación es difícil separar las interpretaciones de los usos. Y la última, y más importante, es que a los validadores se les libera de la responsabilidad de evaluar las consecuencias de los programas de evaluación, delegando esa responsabilidad a otros, lo cuales pueden ser los mismos usuarios a quienes les atañe el impacto del uso de dichas pruebas.

El segundo modelo que propone Kane (2016) es el que utiliza las consecuencias como indicadores, en este se contempla la validación de un programa de evaluación en el que se evalua qué tanto el programa logra los resultados previstos y se revisa detalladamente las posibles consecuencias tanto postivas como negativas de la implementación de dicho programa. Esta opción sigue la propuesta de Messick respecto a la conformación de una estructura que tenga en cuenta las implicaciones sociales de los programas de evaluación y las bases teóricas del constructo de los puntajes de la prueba, de tal manera que si se encuentra que el programa tiene consecuencias adversas se puede ver afectada la validez del programa, ya que la evaluación de las consecuencias permitirian dar cuenta, de manera crítica, el alcance que tienen las interpretaciones no pretendidas. Estas interpretaciones pueden representar vacios o aspectos que nutran a posteriori el constructo.

El tercer y último modelo es el de la interpretación y el uso, hace una distinción entre las interpretaciones de los puntajes y los usos de los puntajes. Para hacer un proceso de validación se deben evaluar qué tanto cada una de las interpretaciones responde a lo planteado inicialmente por el programa de evaluación. Para ello es necesario que se especifiquen claramente las interpretaciones y los usos propuestos de la prueba y así poder evaluarlos a corto y mediano plazo. Respecto a la evaluación de las consecuencias, existe una gama amplia de posibilidades, pero se pueden considerar dos como básicas ya que son evaluado en términos grupales y son ampliamente reconocidos de interés público, estos puntos son: (a) evalauar el impacto diferencial en contra de un grupo en particular (fuentes de sesgo) y (b) efectos sistemáticos no deseados (p. ej. en contextos educativos). Es decir, un validador puede encontrar que las consecuencias están justificadas o no para el uso

propuesto, considerando tanto las consecuencias esperadas como las no esperadas (e.g. impacto diferencias o efectos sistemáticos): Estas estarían justificadas si el programa logra los resultados esperados y no se identifican consecuencias negativas significativas; y no estarían justificadas cuando el programa no logra los resultados esperados y/o tiene efectos negativos significativos en la población.

En conclusión, para Kane (2013, 2016) la aproximación basada en argumentos provee una guía para identificar las debilidades en las interpretaciones y usos propuestos de las pruebas y presentar las implicaciones de la propuesta de evaluación para que sea evaluada su completitud, coherencia y viabilidad basada en la evidencia empiírica recogida y las inferencias hechas a partir desus resultados. Kane (2016) considera que en la evolución en el concepto de validez siempre ha existido la necesidad de justificar las aseveraciones basada en los puntajes de las pruebas.

El consenso sobre el concepto de validez

Newton y Shaw (2016) consideran que lograr un consenso entre los profesionales sobre el concepto de validez es deseable, ya que una definición técnica precisa facilita la comunicación sobre un referente común, contrario a lo que sucede con usos divergentes que dificultan la comunicación. Sin embargo, los autores han encontrado tres tipos de objeciones a la deseabilidad de alcanzar un consenso y a las tres responden con otro argumento respecto a la favorabilidad de lograr un consenso.

La *primera objeción* expresa que no se debería buscar un consenso sobre una definición técnica precisa porque la palabra validez no se admite esa claridad conceptual. Lo mejor que se puede esperar es un consenso implícito sobre la aplicación correcta del término en diferentes contextos. Sin embargo, estos dos autores consideran que precisamente los grandes desacuerdos sobre la mejor forma de emplear la palabra *validez* son incompatibles con la idea de una familia de conceptos relacionados que puedan ser aplicados de forma consistente.

Según la *segunda objeción,* no se necesita buscar un consenso sobre el concepto de validez porque una regulación del uso del término limitaría el desarrollo y el avance de la comprensión del uso de las pruebas pues ésta siempre estará en constante flujo y el

desacuerdo es el motor del progreso en la ciencia. Newton y Shaw (2016) consideran que esta objeción podría ser más robusta si los profundos desacuerdos en el debate sobre el concepto de *validez* se dieran sobre conceptos fundamentales básicos. Sin embargo, en este caso la palabra *validez* es utilizada más bien como una etiqueta para un concepto, es decir que responde a una convención de uso de términos, y no como una verdad. Esto apoyaría la presunción de que buscar un consenso es fundamental.

La *tercera objeción* es que es mejor un consenso porque la *validez* se debe ver como un término controvertido, visto desde la perspectiva de Gallie (1956), y por tanto puede haber un acuerdo general sobre algunos criterios básicos, pero diferentes grupos pueden valorar estos criterios de manera diferente y, lo que es más importante, cada grupo promovería su propio enfoque de evaluación como la verdadera encarnación de dicho concepto. Ante esta objeción, Newton y Shaw argumentan tres razones por las que esta no es viable para el concepto de la *validez.* En primer lugar, el debate sobre lo que debería abarcar la «*validez*» se ha caracterizado por un desacuerdo fundamental sobre la pertinencia de los criterios evaluativos -algunos rechazan la evaluación ética y otros la abarcan-, lo que sugiere que los académicos y demás usuarios no han disputado un concepto único. En segundo lugar, el debate no puede explicarse en términos de diferencias fundamentales entre grupos en términos de sus valores, ya que nadie cuestiona la importancia de la evaluación ética en las pruebas ni la importancia de la evaluación científica. Finalmente, una característica definitoria de conceptos esencialmente controvertidos es que su ambigüedad no puede ser resuelta por mandato, debido a que ningún regulador es universalmente reconocido; sin embargo, se reconoce la notable tenacidad de organizaciones como la AERA, la APA y el NCME para producir sucesivas ediciones de los Estándares en las que precisamente esto es lo que se intenta.

Aunque Newton y Shaw (2016) muestran estas tres objeciones también presentan tres alternativas para lograr un consenso en el concepto de validez, las cuales en general son: eliminar la ambigüedad, aceptar la ambigüedad o retirar la palabra «*validez*» del vocabulario de la evaluación por medio de pruebas.

La *primera opción* es eliminar la ambigüedad y lograr un acuerdo para tener una definición técnica, sin embargo, es un trabajo complejo lograr un consenso principalmente

entre las posiciones liberales y conservadoras, cuyo punto de discordia se basa en las consecuencias. Para los conservadores, obviar las consecuencias de la definición de validez significa correr el riesgo de dejar de lado las consideraciones éticas en el que hacer de la evaluación por medio de pruebas, ya que nadie asumiría la responsabilidad por ellas. Este caso fue estudiado por Cronbach (1988) y Messick (1989), y ellos, al observar casos reales de contextos evaluativos, ilustraron cómo si se excluyen las implicaciones éticas de la palabra *validez*, se puede correr el riesgo de parecer que se está eximiendo a los evaluadores de cualquier responsabilidad de investigar las consecuencias adversas de usar pruebas. Por el contrario, para los conservadores, incluir la palabra «consecuencias» en el concepto de validez es complicarlo aún más, de tal manera que su comprensión puede ser confusa para los usuarios. Se podría correr el riesgo que al querer abarcar todo, se convierta en un concepto inútil en la realidad, que al ser tan amplio no pueda tocar los puntos fundamentales propios de los procedimientos que aseguran una alta calidad en la medición.

Otros de los puntos que resaltan Newton y Shaw, es que la palabra *validez* ha sido elogiada por muchos académicos, «*es **la** palabra, es **nuestra** palabra*» (Newton y Shaw, 2016, p.188), que tiene cimientos emocionales en cada uno, y para la cual se quiere que abarque todo lo posible en cuanto a la buena práctica de evaluación. En esto están de acuerdo con que es muy arriesgado que con ella se abarque todo y que se termine sin tocar nada. Para ellos cuanto, el debate de gran validez del siglo XX (Crocker, 1997) bien pudo haber degenerado en el gran estancamiento de validez del siglo XXI. Así, parece que las perspectivas de llegar a un consenso sobre una definición técnica precisa de validez parecen ser muy bajas.

La *segunda opción* es aceptar la ambigüedad en el uso del término validez, aceptar que en la práctica éste *ya* denota todo lo relacionado con la calidad de una prueba. Este uso resulta aún más extremo que lo buscado por el grupo denominado liberal por Newton y Shaw. En este caso, la palabra *validez* sería el término más importante para el uso de pruebas en educación y psicología, pero sin una definición técnica precisa; simplemente denotaría una evaluación positiva de cualquier aspecto de una prueba. La utilidad de esta opción se encontraría en la comunicación en un nivel que no requiera de precisión entre profesionales o con el público lego.

Un ejemplo exitoso que dan los autores es el del uso del adjetivo saludable. Este último se puede usar apropiadamente en contextos disímiles para dar a entender que se da una evaluación positiva de algún aspecto de la vida (mente saludable, cuerpo saludable, estilo de vida saludable, etc.).

A pesar del atractivo de esta opción, Newton y Shaw señalan que, en tanto que su ejecución conlleva eliminar del significado del término todo excepto los significados más generales, y que es tal la carga emocional que tiene el término en la comunidad de evaluación, puede ser más simple un consenso en renunciar al término validez que uno en el que este término se vacíe tan completamente de significado.

Finalmente, la *tercera opción* es, precisamente, retirar el término validez. La única objeción que Newton y Shaw consideran para esta opción es que no por retirar el término validez se acabaran los desacuerdos sobre los conceptos que se discuten y sobre los que hay fuertes desacuerdos con respecto a la misma. Sin embargo, los autores consideran que esta es la opción más viable para lograr un consenso en el área de evaluación con relación a la validez. En primer lugar, argumentan que la objeción señalada en realidad no afecta que se use o no el término validez puesto que el desacuerdo se da básicamente sobre cómo aplicar adecuadamente la etiqueta de validez, no sobre cómo comprender los conceptos subyacentes de calidad. En segundo lugar, argumentan que el debate actual en torno al término tiene como consecuencias indeseables que fomenta la toma de partidos, con lo cual se dificulta apreciar los aprendizajes ofrecidos por cada perspectiva y se proyecta una idea de que los participantes en el debate tienen posturas totalmente incompatibles con relación a la naturaleza y objetivo de la evaluación en educación y psicología. Por último, argumentan que, a pesar de los desacuerdos, ninguno de los participantes en el debate negaría la importancia de los diversos conceptos subyacentes que están involucrados.

Método Delphi

El método Delphi es una herramienta que se enfoca en maximizar el juicio de los participantes y sus habilidades para la toma de decisiones (McKenna, 1994). Con esta técnica se busca recolectar, de forma sistemática, la opinión de expertos sobre un tema específico con el fin de explorar, comparar juicios y obtener el consenso de sus opiniones.

La información es obtenida por medio de la aplicación de cuestionarios intensivos intercalada con retroalimentación controlada (Dalkey y Helmer, 1963; Reid 1988). Esta técnica ha sido ampliamente utilizada en áreas como las humanidades, la salud y las políticas públicas.

El nombre de la técnica es retomado de la mitología griega (Marchais-Roubelat y Roubelat, 2011). Keenny, Hasson y Mc Kenna (2011) describen el origen de la palabra en el siguiente párrafo:

> «*El nombre de 'Delphi' se deriva del oráculo de Delfos. Delfos [Delphi en inglés] es un sitio arqueológico en Grecia en la cara sur-oeste del monte Parnaso. En la mitología griega, Delfos era la ubicación del oráculo más importante en el mundo griego clásico, y un sitio importante para la adoración del dios Apolo. El dios Apolo llegó a ser el Señor de Delfos después de matar a Pitón, dragón que custodiaba el oráculo, también fue conocida su habilidad para prever el futuro (Linstone, 1978). La leyenda cuenta que las profecías de Apolo se transmitían a través de intermediarias femeninas, conocidas como Pitonisas, nombre derivado de Pitón, fuente de sabiduría en la antigua Grecia (von der Gracht, 2008). Tenía que ser una mujer mayor, de vida intachable, elegida entre las campesinas de la zona. En un estado de trance, inducido por los vapores que surgían de una sima en la roca, la Pitonisa (o sacerdotisa) se sentaba en un trípode sobre una abertura en la tierra y le comunicaban respuestas de Apolo a los sacerdotes y quienes las traducían a los peticionarios. Gentes de todas partes consultaban al oráculo de Delfos en una gama de temas, incluyendo asuntos importantes de la política pública, asuntos personales, el resultado de las guerras y la fundación de colonias. Por lo tanto, el término 'Delphi' se convirtió en sinónimo de recibir un buen juicio sobre un tema*». (p. 2, traducción propia).

Sin embargo, el desarrollo del método Delphi se dio al comienzo de la guerra fría (Custer, Scarcella y Stewart, 1999). Según Kenny et ál. (2011), en el año 1946, la *Douglas Aircraft Company* inició el proyecto RAND con el objeto de analizar y realizar proyecciones respecto al uso de armamento y la capacidad de ataque, para ello uso métodos cuantitativos y cualitativos de investigación; no obstante, los métodos que emplearon no

fueron los más eficaces para responder al tipo de cuestionamientos que se tenían, puesto que los métodos cuantitativos requerían el cumplimiento de parámetros que no se ajustaban al tipo de información con la que contaban, y los métodos cualitativos mostraron tres grandes problemas para la obtención de información útil para el proyecto: la influencia de personalidades dominantes durante los grupos focales u otras técnicas de discusión grupal, el ruido en la información y la presión de grupo que pudieron sentir algunos de los participantes cuando dieron sus opiniones (Dalkey, 1969, p.14).

Durante los años cincuenta en el Proyecto RAND, Olaf Helmer, Norman Dalkey y Nicholas Rescher iniciaron la definición del método Delphi con la premisa de que las predicciones estadísticas individuales eran más fuertes que las predicciones de grupo con encuentros cara a cara puesto que permitían disminuir estas limitaciones (Kaplan, Skogstad y Girshick, 1949; Dalkey, 1969). La primera vez que se utilizó este método fue para obtener la opinión de expertos sobre la probabilidad, frecuencia e intensidad de posibles ataques enemigos y el número de bombas atómicas necesarias para destruir un objetivo particular. Este proceso se repitió varias veces hasta lograr un consenso (Dalkey, 1969).

Linstone y Turoff (1975/2002) definen el método Delphi como «*una estrategia que permite estructurar un proceso de comunicación entre un grupo de expertos para tratar un tema complejo*». Este grupo de expertos es visto como un todo, de tal manera, que se utiliza el juicio de todos ellos para organizar y aclarar sus puntos de vista respecto al tema tratado. Para lograr la comunicación estructurada se proporciona: (a) retroalimentación a las contribuciones individuales, (b) evaluación del juicio del grupo, (c) oportunidad para que los participantes revisen sus puntos de vista, y (d) anonimato para las respuestas individuales.

Según McKenna (1994), dentro de las características clave del método Delphi se encuentran: a) los participantes son un panel de expertos en el tema de estudio, ellos son quienes suministran la información que será analizada; b) las discusiones sobre el tema por tratar no se desarrollan por medio de encuentros cara a cara entre expertos, sino que su participación durante el desarrollo de la investigación es anónima; c) los cuestionarios o entrevistas se aplican siguiendo una secuencia dirigida por los investigadores; d) se busca que emerja de manera sistemática el consenso de opiniones o juicios entre los expertos; e)

las respuestas de los participantes son anónimas; f) se les realizan dos o más rondas de cuestionarios o entrevistas a los expertos y entre cada una de estas se les debe comunicar el análisis de los resultados de la ronda anterior a los participantes para que lo evalúen y contesten el siguiente cuestionario o finalicen la etapa de preguntas.

Las características propias de esta técnica son bastante claras y les permiten a los investigadores utilizarlas en diferentes contextos de investigación, de igual manera, la flexibilidad que muestra la técnica en su aplicación ayuda a definir diferentes tipos o formas de desarrollo de un estudio utilizando el método Delphi (McKenna, 1994; Hasson, Keeney, McKenna, 2000; Keeney et ál., 2011, Rauch, 1979), de allí que se puedan describir diversas formas de implementación de este método. Algunos tipos de Delphi varían según el propósito de la investigación, tal es el caso del Delphi clásico, el de decisión, el de políticas, el argumentativo y el desagregado, mientras que otros varían de acuerdo con el procedimiento que siguen, como en los casos del Delphi modificado, el de tiempo real, el e-Delphi, el en línea, el tecnológico y el híbrido.

En el *Delphi clásico,* el primer cuestionario pregunta, de manera abierta, a los expertos por sus opiniones respecto a un tema específico. Estas respuestas son analizadas y de acuerdo con este análisis se envía un segundo cuestionario en forma de afirmaciones o preguntas cerradas; cada experto califica las afirmaciones o preguntas según su opinión y envía sus respuestas. Después de analizar las respuestas se les da retroalimentación a los expertos sobre los resultados y se les vuelve a enviar otro cuestionario basado en los resultados de la ronda anterior. Al envío de cuestionarios, respuestas y retroalimentación se le denomina ronda. En esta forma de Delphi se hacen de dos a cuatro rondas de preguntas hasta lograr acuerdo en algunos o todos de los ítems, o hasta que la tasa de respuesta decrece.

En el *Delphi de decisión* se realiza el mismo proceso que en el Delphi clásico, la diferencia consiste en que el propósito de esta forma es mirar los posibles escenarios para tomar de decisiones mas no llegar a un consenso.

En el *Delphi de políticas* se busca llegar a un consenso o acuerdo en la definición o implementación de políticas públicas. Se puede realizar como un Delphi clásico o modificado.

El *Delphi argumentativo* se enfoca en la búsqueda de argumentos objetivos y relevantes para implementación de políticas, por lo tanto, proviene del Delphi de políticas y no busca llegar a un consenso entre expertos.

El *Delphi desagregado* no tiene como objetivo lograr un consenso, sino identificar diferentes posibles escenarios a futuro para poder orientar la toma de decisiones. Se realiza por medio de análisis de conglomerados.

En el *Delphi modificado* las primeras preguntas se hacen con grupos focales o entrevistas cara a cara con cada uno de los expertos, mas no con cuestionarios de pregunta abierta. Las siguientes rondas se realizan como el Delphi clásico, sin embargo, se pueden emplear menos rondas que en la versión clásica.

En el *Delphi en tiempo real* se utiliza el mismo proceso de rondas, sin embargo, los expertos pueden estar al tiempo en el mismo sitio de la aplicación, el consenso es alcanzado en tiempo real y no a posteriori; puede ser en una conferencia o en un panel de consenso (Hsieh, Tzeng, Wu, Kao y Lai, 2011).

En el *e-Delphi* y *Delphi en línea* se realiza el mismo proceso del Delphi clásico; sin embargo, los cuestionarios son administrados y diligenciados por e-mail o por encuesta en línea (Hsieh et ál. 2011).

El *Delphi tecnológico* tiene el mismo procedimiento que el Delphi en tiempo real, sino que se realiza por medio de una plataforma que les permite a los expertos contestar a la vez, y recibir retroalimentación de forma automática.

El *Delphi hibrido* combina grupos focales para formular las preguntas de los cuestionarios cerrados, hacer la retroalimentación con las rondas de preguntas cerradas del Delphi clásico, y la interacción de expertos y la evaluación de ideas de la técnica de grupos nominal. Esta técnica facilita la participación, el compromiso y la interacción entre expertos reales en contextos reales generando cambios de opinión más robustos sustentados en análisis de datos cualitativos y cuantitativos (Landeta, Barrutia y Lertxundi, 2011).

Panel de expertos

Lo primero que hay que tener en cuenta para conformar un grupo de expertos es decir qué es un experto, para Kenney et ál. (2011) en el método Delphi no existe una definición

estable sobre qué es un experto, según McKenna, (1994), para algunos, el panel de expertos debe estar conformado por grupo de personas muy bien informado de un tema, para Goodman (1987), por especialistas en un campo, según Davidson, Merritt-Gray, Buchanan y Noel, (1997), Lemmer (1998) y Green et ál., (1999) por personas conocedoras de un tema específico. Para algunos no basta que la persona sepa del tema, sino que debe estar involucrada en el contexto propio de la problemática por estudiar (Keeney et ál., 2011).

El grupo de expertos debe representar una amplia gama de las posiciones sobre el tema por discutir, con diferentes experiencias, derivadas de diferentes áreas o contextos; sin embargo, debido a que cada Delphi es particular, necesario plantear de manera clara los criterios de inclusión antes de invitar a participar a los expertos, ya que de esto depende en gran parte la validez del estudio (Landeta, 2006). Respecto al tamaño del grupo de expertos depende del problema por tratar, por lo general pueden ir de 10 a 200 personas, no obstante, existen estudios Delphi con menos y con más personas de lo que es usual, todo depende del tema y nivel de experticia que sea necesaria para responder la pregunta del estudio Delphi (Keeney, 2011; Skulmoski, Hartman y Krahn, 2007; Powell, 2002).

Respecto al anonimato de los participantes, este no siempre se puede garantizar, primero, porque los investigadores conocen a los expertos y, segundo, porque los expertos se pueden conocer entre ellos, aunque no siempre les será fácil identificar quien es el expositor de alguna idea, por eso se habla de un cuasianonimato (Goodman, 1987, McKenna, 1994). Sin embargo, lo que puede motivar a los expertos a modificar su postura son los resultados del análisis grupal, mas no la presión de la figura que tenga mayor peso respecto a otra, las opiniones tienen igual peso en el análisis.

Procedimiento

En el método Delphi se busca la opinión de los expertos entendida como un juicio que puede o no estar respaldado con evidencia. En busca de obtener un juicio experto, se lleva a cabo un procedimiento específico que se desarrolla a través de rondas de aplicación de cuestionarios de pregunta abierta o cerrada, análisis cualitativos o cuantitativos, construcción de instrumentos e informes de retroalimentación a los expertos. Según Keeney et ál. (2011), Aichholzer (2009), Ortega (2008), Skulmoski et ál. (2007) y Okoli y

Pawlowski (2004), en general, los pasos para realizar un estudio Delphi son: a) identificar el problema; b) determinar la viabilidad del estudio; c) identificar y establecer el panel de expertos; d) definir el criterio de nivel de consenso; e) realizar las rondas de diseño y aplicación de cuestionarios; f) realizar los análisis de los datos pertinentes al tipo de información (cualitativos o cuantitativos); g) identificar los ítems en los que existe consenso y en los que no; y h) realizar los informes de retroalimentación para las dos últimas rondas.

Viabilidad. Teniendo en cuenta la pregunta de investigación y el objetivo particular del Delphi, el primer paso para iniciar el proceso es evaluar la viabilidad de realizar un estudio con las exigencias del estudio Delphi. Según Linstone y Turoff, (1975/2002) para que tenga sentido realizar un estudio Delphi es importante que: a) el problema de investigación no precise un análisis técnico usual de investigación, pero se beneficie del análisis de juicios subjetivos, b) existan posturas encontradas respecto al tema en estudio; c) se busque explorar diferentes opiniones de expertos; d) los expertos se encuentren geográficamente distribuidos; y e) se busque encontrar consenso.

Keeney et ál. (2011), Aichholzer (2009), Skulmoski et ál. (2007), y Okoli y Pawlowski. (2004) afirman que también es importante evaluar la disposición de los recursos necesarios para llevar a cabo una investigación en la que se debe tener comunicación fluida con los expertos, se deben realizar análisis de datos cualitativos y cuantitativos, e invertir en licencias para plataformas de encuestas y de programas de análisis de datos, entre otros.

Panel de expertos. Cuando ya se tiene claro el problema y se ha determinado que un estudio Delphi es la estrategia pertinente para su resolución, se proponen los expertos que pueden conformar el panel. Algunas recomendaciones genéricas que se pueden tener en cuenta para elegir al experto son que tengan: a) conocimiento o experiencia en el tema que se está trabajando, b) capacidad y disposición para participar c) certeza que cuenta con el tiempo para responder los cuestionarios, y d) buenas habilidades de comunicación escrita (Ziglio, 1996; Skulmoski et ál., 2007). Luego de que se han determinado quiénes pueden ser los expertos, se establece el tamaño del panel, se obtienen sus datos de contacto y se desarrollan las estrategias de comunicación y de administración de los cuestionarios. Después se contacta a todos y cada uno de los expertos y se les invita a participar en el

estudio, se detalla el objetivo de la investigación, se describe el procedimiento, precisando que son varias rondas y que se utilizan diferentes tipos de cuestionarios, y el tiempo que conlleva, el cual puede tomar varios meses. Es relevante tener una comunicación continua y amena con los expertos, pues de esta depende en gran parte el compromiso y entusiasmo que los expertos tengan con el estudio. También se debe ser muy claro en notificar las fechas en que les llegarán los cuestionarios y el tiempo con el que cuentan para responder (Skulmoski et ál., 2007; Keeney et ál., 2011; Powell, 2003).

Consenso. En el estudio Delphi se habla del *acuerdo*, la *estabilidad* y el *consenso*, estos tres conceptos son importantes a la hora de establecer el consenso, ya que utilizan diferentes estimaciones, tienen diversas interpretaciones y, en el momento del análisis, pueden dar pie a confusiones, por lo que es necesario tener clara su definición y su uso (von der Gracht, 2012; Keeney, 2011). El «*acuerdo*» se entiende como el grado en el que los expertos están a favor de una afirmación (Keeney et ál., 2011). Dajani, Sincoff y Talley (1979) definen la «*estabilidad*» como «la consistencia de las respuestas entre rondas sucesivas en un estudio» (p. 84), también afirman que para que el acuerdo sea significativo es necesario que se haya alcanzado la estabilidad en las respuestas, por lo que esta se convierte en un criterio importante para decidir el momento en el que se detienen las rondas y se estima el consenso en un Delphi. La estabilidad se puede estimar tanto intragrupo como intraindividuo (von der Gracht, 2012).

Por su parte, si bien el «*consenso*» es uno de los pilares de un estudio Delphi es un término polémico y la forma de su medición varía mucho (von der Gracht, 2012; Powell, 2003). Keeney et ál. (2011) lo define como el grado en el que los expertos están de acuerdo unos con otros en relación con una afirmación, Mitchell (1991) describe que este puede ser definido como una opinión grupal, un acuerdo general o una solidaridad grupal en sentimientos y creencias. De igual forma, la manera para establecer el criterio de consenso tampoco se ha definido de manera rigurosa (Mitchell, 1991). En este caso el consenso se puede definir como un acuerdo colectivo, que surge de la reunión y trabajo colaborativo entre las partes interesadas; usualmente este trabajo se realiza con un facilitador hasta que se alcanza uno o varios puntos de convergencia. En el caso del estudio Delphi, el facilitador

es el investigador y el acuerdo se busca mediante el uso de rondas de cuestionarios (Keeney et ál., 2011).

Keeney et ál. (2001, 2006, 2011) insisten en que encontrar consenso en el Delphi no significa que se hayan encontrado respuestas correctas respecto al tema, tampoco reemplaza las revisiones científicas a profundidad que se realizan en investigaciones originales o en artículos producto de estas. Si bien el método puede ayudar a revisar puntos de vista y ganar algo de consenso, no se pude tomar como algo definitivo, sino como una técnica útil en el proceso de llegar a un acuerdo o el estudio integral de ciertas temáticas (Keeney et ál., 2006). El nivel de acuerdo para establecer consenso lo determina el equipo investigador antes de aplicar estos cuestionarios de las rondas dos y tres; es importante tener en cuenta que si se define un criterio muy exigente es posible que no se llegue a un consenso (Kenney et al, 2011; von der Gracht, 2012).

No todos los estudios Delphi buscan llegar a un acuerdo, los estudio Delphi de políticas, el argumentativo y el de decisiones buscan, por el contrario, enfocan su atención en los desacuerdos y a describir como son las distribuciones de esas diferencias, y así comprender a profundidad los puntos de vista de los expertos, que pueden ser personas involucradas en una misma problemática, pero desde posiciones muy diversas o incluso contrarias (von der Gracht, 2012).

Rondas. El proceso de recolección de información se realiza a través rondas de aplicación de cuestionarios. En el estudio Delphi, el promedio de rondas utilizados es de tres; se ha visto en varios casos, que dos o tres rondas tienden a ser suficientes para recolectar la información. También se ha observado que a medida que aumentan las rondas de cuestionarios, la tasa de respuesta por parte de los expertos disminuye. Sin embargo, es importante anotar que si las posiciones de los expertos son muy dispares se podrían necesitar más rondas (Skulmoski et ál., 2007; Keeney et ál., 2011). A continuación, se describen los pasos por seguir en cada una de las rondas.

Primera ronda. El objetivo de la primera ronda es generar ideas y se les pide a los expertos que comenten acerca del tema en cuestión. Por lo tanto, se emplean cuestionarios de pregunta abierta, que les permiten a los expertos expresarse con libertad respecto al tema

en estudio. En esta ronda también se recoge información sociodemográfica de los panelistas, para describir las características del panel y, especialmente, para confirmar los datos que les dan crédito de su experticia, por ejemplo, edad, afiliación, profesión, años de experiencia y publicaciones (Keeney et ál., 2011).

Para la construcción del cuestionario, se recomienda tener como mínimo cinco ítems y como máximo diez, esta es una ronda cualitativa y se debe buscar equilibrio entre la cantidad de preguntas y la pertinencia y relevancia de estas respecto al tema estudiado. Es importante tener en cuenta que estas respuestas dependen la cantidad de preguntas de la siguiente ronda, y se debe asegurar tener una tasa de respuesta razonable por parte de los expertos. Las respuestas abiertas pueden dar lugar a cuestionarios muy largos para la siguiente etapa si se incluyen todas las opiniones de todos los expertos. En este caso, se corre el riesgo de no poder mantener el grupo de expertos a lo largo del estudio (Keeney et ál., 2011).

Por otra parte, también es importante monitorear la finalización, el envío y la recepción del cuestionario, para ello es necesario estar en continua comunicación con los expertos y atender sus dudas de manera diligente para motivar y reforzar su participación (Keeney et ál., 2011).

Cuando se han recibido las respuestas de los expertos, se inicia el proceso análisis cualitativo de esta información, es decir, se establecen las categorías de análisis, se organiza la información, se analiza y se generan las premisas o afirmaciones que constituyen el segundo cuestionario.

Segunda, tercera o cuarte ronda. En las siguientes rondas, se utilizan cuestionarios estructurados con los ítems generados en la primera ronda y se incorpora la retroalimentación de los expertos. Una de las ventajas del estudio Delphi es que logra comprometer y motivar al panel de expertos puesto que ellos se ven involucrados en el desarrollo del instrumento y crean un sentido de pertenencia y aceptación respecto a los hallazgos (McKenna, 1994).

El cuestionario está compuesto por afirmaciones que responden a las categorías trabajadas en la primera ronda, se recomienda que el instrumento sea extenso y que tenga las afirmaciones organizadas por área temática para que el trabajo sea más cómodo para los

expertos. Los miembros del panel de expertos deben calificar cada uno de los ítems o prioridades en una escala pertinente. Esto podría ser una escala de Likert de siete puntos de «muy importante» a «sin importancia» o una escala de cinco puntos de «muy de acuerdo» a «fuertemente en desacuerdo» (Keeney et ál., 2011).

A los miembros del panel se les envía una comunicación en la que se les presenta la nueva ronda, se les dan instrucciones y se les anexa el cuestionario. Es importante que quede claro el tiempo estipulado para diligenciarlo y la fecha límite de envío de vuelta. El Delphi permite una rápida recolección de datos si la retroalimentación es controlada.

Es deseable que los miembros del panel de expertos permanezcan hasta el final del proceso para lograr el consenso. Para aumentar la tasa de respuesta de los expertos es importante que ellos se den cuenta y sientan que sus compañeros están interesados en la discusión, y enviar recordatorios de que con base en sus respuestas se construyen los cuestionarios de cada ronda (Buck, Gross, Hakim y Weinblatt, 1993).

Al igual que en la primera ronda, los recordatorios de seguimiento deben enviárseles a los miembros del grupo de expertos cuando sea necesario. Como la segunda ronda puede ser un momento en el que muchos miembros de un panel de expertos abandonen, se debe hacer todo lo posible para mantener su interés en el estudio (Linstone y Turoff, 1975/2002).

Uno de los problemas que se pueden presentar en el estudio Delphi es una disminución en la tasa de respuesta; las tasas de respuesta pobres son una característica de la ronda final del Delphi. Esta ha sido una crítica fuerte a esta metodología y por eso se hacen dos o tres rondas en vez de cuatro. Hacer entrevistas cara a cara en la primera ronda ha demostrado que logra que los expertos se involucren hasta el final del estudio (McKenna, 1994).

Análisis de datos y resultados

El análisis de la información en un estudio Delphi se realiza tanto a nivel cualitativo como cuantitativo. Respecto al análisis cualitativo, en la primera ronda se crea un instrumento de preguntas abiertas que estimula que los expertos expresen sus opiniones de manera libre, por lo que se espera realizar un análisis de contenido de esta información con el objeto de crear las afirmaciones para construir el segundo cuestionario. En cuanto al análisis cuantitativo, éste se basa en los datos recogidos en los dos siguientes instrumentos

que son de pregunta cerrada con escalas discretas, ordinales o nominales. Teniendo en cuenta este tipo de información, se realizan análisis estadísticos que describen la tendencia de las respuestas y el nivel de consenso de los expertos.

Análisis cualitativo. El objetivo del análisis cualitativo en un estudio Delphi es generar grupos de afirmaciones por temas similares de acuerdo con las opiniones dadas por los expertos. Usualmente, se utiliza el análisis de contenido como técnica para realizar esta parte del estudio, cualquier tipo de análisis de contenido es útil, sin embargo, un análisis simple responde adecuadamente a las necesidades de esta fase del estudio (Keeney et ál., 2011).

El análisis de contenido es «una técnica de interpretación de textos, ya sean escritos, grabados, pintados, filmados..., u otra forma diferente donde puedan existir toda clase de registros de datos, trascripción de entrevistas, discursos, protocolos de observación, documentos, videos» (Andreu, 2002; p. 2). Esta técnica se basa en la lectura del texto siguiendo el método científico, por lo que se debe procurar que sea una lectura sistemática, objetiva, replicable y válida (Andreu, 2002).

El contenido del texto puede ser interpretado de una forma directa y expresa, o de forma mucho más semántica, buscando el sentido latente en el texto. En un análisis temático en el que se estudia la presencia de conceptos o palabras, sin considerar relaciones, solo se centra en el contenido expreso (Escalante, 2009). Tanto los datos expresos como los latentes tienen significado y cobran su importancia ubicados en un contexto particular, por lo tanto, el contexto es el marco de referencia del que debe partir el investigador para realizar el análisis, ya que en este están inmersos los participantes, brinda toda la información que se puede conocer con anterioridad y da la guía para realizar las inferencias a partir del texto y, con ello, entender el significado de lo que se está expresando (Andreu, 2002).

Para el caso del estudio Delphi, se realiza un análisis del contenido expreso del texto, Andreu (2002) propuso los siguientes pasos para llevar a cabo este tipo de análisis: a) determinar el objeto o tema de análisis; b) determinar las reglas de codificación; c) determinar el sistema de categorías; d) comprobar la fiabilidad del sistema de codificación-categorización y; e) plantear las inferencias.

Análisis cuantitativo. En la segunda y tercera ronda se realizan análisis cuantitativos a las respuestas dadas por los expertos en cada uno de los cuestionarios. En estas etapas se busca identificar el nivel de acuerdo con cada una de las afirmaciones, la estabilidad de las respuestas para cada experto o del grupo y la presencia o no de consenso para cada una de las afirmaciones del cuestionario.

Para lograr identificar el nivel de *acuerdo* que tienen los expertos con cada afirmación se realizan análisis descriptivos de su tendencia central. Keeney et ál. (2011) muestran un resumen de los estadísticos reportados en diferentes estudios Delphi, en este se puede ver que los estadísticos más usados para describir el acuerdo son porcentajes, la mediana, la media o la moda.

Respecto a la *estabilidad*, von der Gracht (2012) señala que es común que en los estudios Delphi se establezca un criterio para decidir cuándo detener las rondas. Algunos de los investigadores suelen tomar esta decisión antes de empezar la aplicación, sin tener en cuenta un criterio estadístico de estabilidad o de consenso, sino basándose en un análisis costo beneficio, en el que consideran el presupuesto, el tiempo, las particularidades del panel de expertos, o porque suponen que con otra ronda no se obtendrá mayor cantidad de información de la ya recogida. En otros estudios Delphi se han utilizados estadísticos para describir el cambio en las respuestas de los expertos entre rondas sucesivas y así ayudar a decidir cuándo detener las rondas de preguntas (Dajani et ál., 1979; Chaffin y Talley, 1980) Entre los métodos estadísticos reportados en la revisión hecha por von der Gracht (2012) aparecen el Chi cuadrado de independencia, la prueba de cambio McNemar, la prueba Wilcoxon, el coeficiente de correlación intraclase Kappa, el coeficiente de correlación Spearman, el coeficiente de concordancia W de Kendall pruebas t y pruebas F.

En relación con el *consenso,* Keeney et ál. (2011) afirman que para establecer el criterio de consenso se depende de la escala, en el caso de los ítems de tipo Likert se determina un porcentaje en una categoría que supere el criterio establecido por los investigadores, este puede ser mayor de 65 %, 70 % o 80 %. También señalan que los estadísticos más usados para realizar la retroalimentación a los expertos son la mediana y el rango intercuartil (RI) o la media y la desviación estándar. Pocos estudios han reportado el uso de la correlación, los coeficientes de Kruskall Wallis y la W de Kendall, y el Chi cuadrado.

Otros de los índices que se han utilizado para estimar el consenso en estudios Delphi son nombrados por Birko, Dove y Özdemir (2015), los cuales realizaron una evaluación de nueve índices y su dependencia respecto al tamaño del panel de expertos, cantidad de preguntas y qué tanto los expertos cambian de opinión. Los nueve índices fueron el índice de Moivre, el acuerdo dos a dos expertos, el acuerdo dos a dos agrupado, la versión de extremos del acuerdo dos a dos agrupado, el Kappa de Fleiss, la frecuencia modal, la frecuencia del intervalo modal y el rango intercuartil (RI).

Por su parte von der Gracht (2102), en la revisión que hace sobre medidas de consenso en estudios Delphi, señala que usualmente para determinar el consenso se utilizan criterios subjetivos y estadísticos descriptivos; aunque muchas veces tanto el criterio como la cuantificación de su grado es escogida arbitrariamente. También presenta diversos tipos de estadísticos desde medidas de asociación hasta medidas de tendencia central y dispersión. Los métodos descriptivos que nombra son el sistema de votación, el promedio del porcentaje de la mayoría de las opiniones, las medidas de tendencia central (media, mediana, moda) analizadas junto con las medidas de dispersión (rango, desviación estándar, rango intercuartil y coeficiente de variación). En su revisión von der Gracht es enfático que es importante tener en cuenta el tipo de escala utilizada en el cuestionario para escoger los estadísticos que se van a utilizar para el análisis.

En otros estudios, mucho más recientes, utilizan ítems con escalas difusas (fuzzy) y realizan análisis propios para este tipo de datos, estos son el caso de las investigaciones hechas por Pankratova y Malafeeva (2012); Ma, Shao y Ye (2011); García y Lazzari, (2000); y Bravo y Arrieta, (2005).

Aplicaciones del método Delphi

Landeta (2006) y von der Gracht (2012) describen cómo durante los últimos treinta años, se ha presentado una creciente tendencia en el empleo del método Delphi como estrategia para la investigación tanto en publicaciones en artículos científicos como en textos. Algunos de los campos en los que se ha utilizado este método son en educación (Bravo y Arrieta, 2005), psicometría (Cruz y Martínez, 2012), medicina (Weir, Hölmich, Schahe, Delahunt y de Vos, 2015), psiquiatría (Jorm, 2015; Jeste, Ardelt, Blazer, Kraemer,

Vaillant y Meeks, 2010), economía y finanzas (Kauko y Palmroos, 2014; Campos, Melían y Sanchis, 2014), mercadeo (Bonnemaizon., Cova, Louyot, 2007), ingeniería (Ma et ál. 2011; Elmer, Seifert, Kreibich y Thieken, 2010) y estudio de políticas públicas (Wentholt, Rowe, König, Marvin y Frewer 2009).

Teniendo en cuenta las características descritas de la metodología de un estudio Delphi, se concluyó que es este tipo de diseño permite responder los objetivos de este proyecto el cual es identificar, entre un grupo de expertos, los puntos en los que existe consenso y aquellos en los que no sobre el concepto de validez en educación y psicología, con el fin de aportar elementos que contribuyan a dar claridad y consistencia a la conceptualización de la validez. De tal manera que se realizó un análisis de las opiniones de algunos de los expertos que han orientado la discusión del concepto de *validez* mediante el uso del Método Delphi en línea.

Método

Participantes

Siguiendo la clasificación de las posturas respecto al concepto de la *validez,* propuesta por Newton y Shaw (2016), se invitaron expertos que representaran, por su contribución teórica y metodológica, las posturas tradicional, conservadora, ultraconservadora y liberal, con el fin de tener la gama más amplia posible de las opiniones investigadas en el campo de la medición y evaluación en la educación y la psicología. Se invitaron 11 expertos, de los cuales siete manifestaron su capacidad y disposición para participar en el estudio Delphi.

Los participantes fueron siete expertos académicos reconocidos que cumplieron con los criterios para conformar el panel de expertos. Estos criterios fueron que durante la última década y media (2000-2016) hubieran: (a) estudiado a profundidad el concepto de *validez*, (b) publicado sus análisis con relación a este concepto en el área de medición y evaluación en psicología y educación en textos especializados y revistas científicas de alto impacto, y (c) liderado la discusión sobre el concepto de *validez* en las últimas décadas tanto en publicaciones de alto nivel como en reuniones académicas. Estos expertos también cuentan con experiencia a nivel académico como profesional en la evaluación de validez de pruebas académicas de alto reconocimiento en Europa y Estados Unidos. Los siete expertos que participaron en el estudio fueron: Denny Borsboom, Gregory Cizek, Michael Kane, Robert Mislevy, José Muñiz, Paul Newton y Stuart Shaw. En la primera ronda participaron siete expertos; en la segunda ronda, seis; y en la última ronda, cuatro.

El doctor Denny Borsboom es profesor de principios de psicología y de psicometría en la Universidad de Amsterdam de la Facultad de Ciencias Sociales y del Comportamiento. Ha publicado varios articulos sobre el concepto de validez, medición, psicopatología y ha publicado dos libros, el primero Measuring the mind: Conceptual issues in contemporary psychometrics publicado por Cambridge University Press, en 2005 y el segundo, Frontiers of test validity theory: Measurement, causation, and meaning, publicado por Routledge en 2013. El profesor Borsboom es miembro de varios comités editoriales de revistas académicas tales como *Frontiers of Quantitative Psychology*, *Educational Measurement: Issues and Practice* and the *European Journal of Personality*.

El doctor Gregory Cizek es profesor de psicometría aplicada, estadística y programas de evaluación y métododos de investigación de University of North Carolina - Chapel Hill. Tiene más de veinte años de experiencia como docente e investigador. Ha publicado varios artículos sobre validez y expuesto sus investigaciones en multiples encuentros académicos sobre evaluación en psicología y educación.

El doctor Michael Kane ocupa la silla Samuel J. Messick en Validez de intrumentos desde el 2009, anteriormente fue Director de Investigación del National Council of Bar Examiners y vicepresidente de investigación American College Testing. Es quien lidera la investigación de la teoría de validez en Educational Testing Service. Ha publicado sus trabajos de investigación en multiples revistas y ha liderado procesos de validación robustos con diversas pruebas en Estados Unidos. El doctor Kane tiene más de 50 años de experiencia en el campo de la psicometría y de la evaluación educativa y psicológica, tanto a nivel académico como aplicado.

El doctor Robert Mislevy ocupa la silla Frederic M. Lord en Medición y Estadística del Educational Testing Service desde 2010. Previo a este cargo fue docente del departamento de Medición, Estadística y Evaluación de la Universidad de Maryland. Ha publicado sus trabajos sobre psicometría y medición en diversas revistas. El doctor Mislevy lleva más de 30 años trabajando en el campo de la psicometría y de la evaluación en educación y psicología.

El doctor José Muñiz Fernández es docente de Psicometría de la Universidad de Oviedo. Ha publicado sus multiples trabajos investigativos de psicometría y evaluación psicológica en diversas revistas, de igualmente ha publicado libros de *Teoría clásica de los test, Teoría de respuesta a los ítems, Psicometría y Análisis de los ítems.* Es miembro de varios comités editoriales de revistas académicas, es director de la revista Psicothema y presidente de la European Association of Methodology. El profesor Muñiz tiene más de 30 años de experiencia en el campo de la psicometría y de la evaluación educativa y psicológica, a nivel académico y aplicado.

El doctor Paul Newton ocupa la silla de investigación en Ofqual. Fue profesor de Evaluación Educativa en el Instituo de Educación en University of London. Ha trabajado como asesor de investigación en diversas entidades de evaluación inglesas como

Associated Examining Board, National Foundation for Educational Research, Qualifications and Curriculum Authority y Cambridge Assessment. El doctor Newton es miembro del comité editorial de la revista *Assessment in Education: Policy, Principles & Practice*. Ha publicado sus trabajos de investigación sobre sistemas de evaluación educativa a gran escala y teoría de validez tanto en revistas como en libros.

El doctor Stuart Shaw es el líder el equipo de investigación del área de evaluaciones internacionales de Cambridge Assessment. Trabaja hace más de 16 años en esta institución y su objetivo es demostrar la validez de las evaluaciones internacionales diseñas y aplicadas por Cambridge Assessment. Su principal área de interés es la evaluación del inglés como segunda lengua, investiga sobre diseños, revisión, implementación de estrategias, nuevos escenarios y análisis de buenas prácticas de evaluación. Ha publicado numerosos artículos en relación con estos temas.

Instrumentos

A lo largo de la investigación se diseñaron, aplicaron y analizaron tres cuestionarios, el formato del primero fue de pregunta abierta y responde a la primera ronda del estudio Delphi; los otros dos fueron de pregunta cerrada tal como se indica para la segunda y tercera ronda. Las rondas del estudio se describen más adelante en este documento.

Primer cuestionario - «Talking about the Concept of Validity». El propósito de este cuestionario fue generar ideas sobre algunos de los temas que se han revisado en la discusión del concepto de *validez*, para ello se plantearon seis preguntas abiertas para que los expertos expresaran sus opiniones respecto a cada una de ellas. Las preguntas indagan sobre la definición de la *validez*, el objeto que se valida, el concepto de validez que presenta los Estándares (2014), las dificultades para lograr un consenso, cómo superar las dificultades nombradas por el mismo experto y algunas propuestas para llevar a cabo el proceso de validación.

El cuestionario inicia con preguntas de identificación del experto (nombre, filiación, país de residencia y correo electrónico) y sigue con las seis preguntas abiertas sobre el concepto de *validez*. El instrumento se diseñó en la plataforma surveygizmo.com, la cual maneja cuestionarios en línea (ver apéndice A).

Segundo cuestionario – «Agreement about the Concept of Validity». El propósito de este cuestionario fue que cada experto indicara su nivel de acuerdo con cada una de las afirmaciones presentadas. Estas frases se derivaron del análisis de contenido de las respuestas dadas por los expertos en el primer cuestionario. Teniendo en cuenta este análisis, las categorías en las que se organizó y diseñó el instrumento fueron: el concepto de validez (16), los Estándares (2014) (14), el consenso sobre el concepto de validez (26), y la validación (18), de tal manera que el cuestionario contó con un total de 74 preguntas cerradas. Estos ítems se califican por medio de una escala discreta, en las que el experto debe graduar su nivel de acuerdo en una escala de 0 a 10, siendo 0 para nada de acuerdo y 10 muy de acuerdo (ver apéndice B).

Tercer cuestionario - «Agreement about the Concept of Validity-Final Part». El objetivo de este cuestionario era que los expertos señalaran su nivel de acuerdo con cada una de las afirmaciones. Las frases que conforman este instrumento son aquellas en las que en los análisis para evaluar el consenso presentaron un $Rango\ Intercuartíl \leq 3{,}0$. El cuestionario contiene 38 afirmaciones, de la cuales 11 son sobre el concepto de validez, 7 son sobre los Estándares (2014), 12 son sobre el consenso sobre el concepto de validez y 8 son sobre validación. Estos ítems se califican por medio de una escala discreta, en las que el experto debe graduar su nivel de acuerdo en una escala de 0 a 10, siendo 0 para «nada de acuerdo» y 10 «muy de acuerdo». El instrumento se diseñó en la plataforma surveygizmo.com (ver apéndice C).

Procedimiento

Viabilidad. El método Delphi se escogió como la metodología más pertinente para identificar, entre un grupo de expertos, los puntos en los que existe consenso y en los que no sobre el concepto de *validez* en educación y psicología. Los criterios para escoger esta metodología se sustentaron en que a) debido a que el problema de la conceptualización de la *validez* ha sido trabajado en varios trabajos teóricos por varias décadas, se han identificado diferentes posturas conceptuales, las cuales son muy diferentes e incluso encontradas; b) si bien los expertos han expresado sus opiniones sobre el concepto de validez, existen temas en los que solo algunos expertos han opinado y falta recoger las

ideas de otros expertos al respecto; por lo tanto era necesario explorar las diferentes opiniones sobre estos temas; c) permite entablar una comunicación entre los expertos que se encuentran geográficamente distantes, los expertos que trabajan sobre el concepto de *validez* se encuentran en diferentes ciudades de Europa o de Estados Unidos; d) permite identificar los temas en los que se presenta consenso y en los que no respecto a un tema determinado; y por último, e) facilita la libre expresión de opiniones gracias al anonimato de los expertos.

Después se evaluó la disponibilidad de recursos para realizar el estudio, para ello se conformó un grupo de investigadores para asesorar y apoyar la revisión técnica de los instrumentos, de los análisis de la información y la toma de decisiones. Este grupo estuvo compuesto por dos estudiantes de maestría en Psicología de la Universidad Nacional de Colombia (en adelante UN), una estudiante de doctorado en Psicología de la UN y un estudiante de doctorado de Purdue University (en adelante, PU), la directora de este proyecto, docente asociada del Departamento de Psicología de la UN, y la codirectora de este proyecto, docente asistente de la Facultad de Educación de PU. Teniendo en cuenta la disponibilidad de recursos y observando la viabilidad para llevar a cabo el estudio Delphi, se inició con el proceso con la consecución de expertos.

Panel de expertos. Para escoger los expertos se plantearon los criterios para definir un experto en el tema, los cuales se describen en el apartado de participantes. Luego se estableció contacto por medio de correo electrónico, en el que se invitaron a participar a diez expertos en el estudio Delphi, de los cuáles siete respondieron a la invitación y con ellos se inició el proceso. Después de su respuesta afirmativa, se les dio a conocer el objetivo de la investigación, el procedimiento de las rondas de cuestionarios propias y la retroalimentación de resultados, los alcances y limitaciones y el tiempo que duraría. El contacto ocurrió con algunos meses de anterioridad a la aplicación de primer cuestionario, con el fin de crear el espacio en sus agendas para responder los instrumentos.

Consenso. Teniendo en cuenta la literatura, en un estudio Delphi se deben definir los criterios para describir el acuerdo, la estabilidad y el consenso. Respecto al *acuerdo*, teniendo en cuenta la amplitud de la escala y la cantidad de expertos, se decidió que la mediana de las puntuaciones fuera el estimador para describir el nivel de acuerdo del grupo,

respecto a cada una de las afirmaciones del cuestionario; y que su interpretación fuera definida así: de 0 a 4 como no de acuerdo; de 4 a 6 como indeciso; y de 7 a 10 como de acuerdo.

En relación con la *estabilidad*, se decidió que no se establecería un criterio estadístico para detener las rondas de cuestionarios, sino que de antemano se definió que con tres rondas se lograba alcanzar el objetivo del estudio, ya que se esperaba que no se llegaría a un consenso total en todas las ideas planteadas, ni que se presentaría un cambio extremo en las opiniones de los expertos. Sin embargo, tras la tercera ronda sí se estimó la homogeneidad de las respuestas entre las dos últimas rondas por medio del rango intercuartil relativo (RIR), ya que este estima el grado de convergencia de opiniones del grupo (von der Gracht, 2012).

Por último, para determinar el *consenso,* se escogió el RI como el estimador para ver la diferencia entre las puntuaciones entre los expertos, este criterio se estableció basándose en la cantidad de expertos, la amplitud de escala y la revisión de literatura respecto a los índices de consenso, particularmente el trabajo de von der Gracht (2012) en el que afirma que el RI es uno de los estimadores más usados en los estudio Delphi y es generalmente aceptado como una forma objetiva y rigurosa para determinar consenso. La pareja de estimadores, mediana y RI, se seleccionó teniendo en cuenta que estos estadísticos son generalmente robustos (Murphy et ál., 1998). Adicionalmente, se consideraron los criterios usualmente reportados para definir mediante el RI que se observa un consenso entre los expertos con respecto a una afirmación (igual o menor a 1 para escalas de 4 o 5 opciones, igual o menor a 2 para escalas con 10 opciones), y se optó por un criterio de 3.0 para el RI; este criterio, algo más amplio que el utilizado en otros estudios, se tomó por cuanto la escala usada en este estudio emplea aún más opciones que otros estudios y porque dado el reducido número de expertos elegibles para participar en el estudio, se esperaba una mayor dispersión de los resultados.

Rondas. La aplicación de cuestionarios en un estudio Delphi se desarrolló por medio de tres rondas sucesivas de cuestionarios.

En la *primera ronda* se diseñó un instrumento que permitiera generar ideas entre los expertos sobre los temas revisados en la revisión teórica. Los temas propuestos al panel de

expertos fueron escogidos por parte del equipo de investigadores, basándose principalmente en la discusión sobre el concepto de *validez* expuesta en los artículos publicados por diferentes académicos durante los años 2014 a 2016: Newton y Shaw, 2014; Padilla e Ibañez., 2014; Sireci, 2014; Borsboom, 2016; Cizek, 2016a; Cizek, 2016b; Kane, 2016a; Kane; 2016b; Markus, 2016; Moss, 2016; Newton y Baird, 2016; Newton y Shaw, 2016a; Newton y Shaw, 2016b; Shepard, 2016; Sireci, 2016a; Sireci, 2016b; Zumbo y Hubley, 2016.

Después de diseñadas y construidas las preguntas, con el grupo asesor se revisaron la pertinencia, la relevancia y la construcción gramatical en inglés de las instrucciones, las preguntas y la plataforma en línea (surveygizmo.com). También se piloteó el acceso a la plataforma y el funcionamiento del instrumento. Cuando ya se aprobó el instrumento, se envió a cada experto el correo electrónico de invitación para diligenciar el cuestionario, en este también se detallaron las condiciones del estudio, las instrucciones del cuestionario, las fechas de acceso al cuestionario y el link de enlace del cuestionario.

El tiempo de diligenciamiento fue de 15 días, durante este periodo se llevaron a cabo comunicaciones con los expertos recordando la importancia de su participación y las fechas límite. Todos los siete expertos contestaron este cuestionario en el tiempo establecido.

Luego de recibir las respuestas dadas por los expertos, se inició el proceso de depuración y organización de la base de datos. Después de organizada la información, se realizó un análisis de contenido sencillo atendiendo a las categorías teóricas inicialmente planteadas con el grupo de asesores y las recomendaciones propias del procedmiento de un estudio Delphi. Las categorías se describen en la tabla 1. Este análisis se llevó a cabo con NVivo versión 11.4.0 (2016).

Tabla 1

Descripción de las categorías teóricas iniciales para el análisis de contenido

Categoría	Definición
Concepto de validez	Describe qué postura tiene al definir el concepto de *validez*, clasificación de posturas según Newton y Shaw (2016).
Objeto de validación	Describe qué es lo que se valida, si el instrumento, las interpretaciones, el procedimiento o aspectos combinados.
Estándares (2014)	Describe el nivel de acuerdo con el concepto de *validez* planteado en los Estándares (2014) y las razones que justifican su opinión.
Consenso	Describe el nivel favorabilidad respecto a llegar a una definición consensuada de la *validez* y detalla los argumentos sobre si está de acuerdo o no con este. También se centra en identificar las estrategias que propone para llegar al consenso, estas pueden ser académicas, políticas o comerciales.
Validación	Describe si propone un procedimiento, un modelo o un "tipo de validación" que debe ser considerada como necesaria o suficiente respecto al proceso de validación.

Siguiendo las recomendaciones para el anális cualitativo de los datos dadas por Keeney et ál., (2011), se continuó con la segmentaron los textos, luego se buscaron similitudes o diferencias sobre cada categoría establecida y se crearon varias subcategorías emergentes y se descartaron subcategorías teóricas planteadas pre-aplicación que no aparecieron en los textos de los expertos. Seguido de esta codificación final, la cual se constituyó la estructura del segundo cuestionario, se extrajeron las frases textuales expresadas por los expertos, para conformar las afirmaciones del instrumento. Después se revisó el estilo de cada frase para que cada una correspondiera a la idea expresada por el experto, fuera clara, independiente de otros textos, y gramaticalmente correcta en inglés. Cada uno de estos procedimientos se realizaron en conjunto por parte de tres miembros del equipo asesor de investigadores del Delphi y se revisaron de manera independiente con cada una de las directoras del proyecto Luego se evaluaron la pertinencia, la relevancia y la forma en inglés de las instrucciones y las afirmaciones, después se piloteó el acceso a la plataforma y el funcionamiento del instrumento. En cuanto el instrumento estuvo aprobado se inició con la segunda ronda.

En la *segunda ronda,* antes de iniciar con la aplicación del instrumento se definieron los criterios de acuerdo, estabilidad y consenso para la segunda y tercera ronda. Después de definir estos criterios de consenso y tener el instrumento listo para la aplicación, se enviaron a los expertos las invitaciones para participar en el segundo cuestionario, el link de acceso al cuestionario, las instrucciones de diligenciamiento y las fechas límite para responder el instrumento. El tiempo que tuvieron los expertos para diligenciar el cuestionario fue de diez días. Seis de los siete expertos contestaron el cuestionario durante este tiempo, el otro experto no respondió el instrumento, no se pudo establecer contacto con él a pesar de tratar de comunicarse con él mediante varios correos electrónicos, por lo que no se conoció la causa de su deserción.

Luego de recibir las respuestas de los expertos, se organizó y depuró la base de datos para iniciar el análisis de datos. El análisis estadístico fue con estimación de los descriptivos básicos, es decir, el mínimo, máximo, mediana, cuartiles y el rango para cada uno de las afirmaciones. Teniendo en cuenta estos resultados, se escogieron las afirmaciones en las que no se presentó consenso para conformar el tercer cuestionario, se diseñó el cuestionario en la plataforma y se evaluaron la pertinencia, la relevancia y la forma en inglés de las instrucciones y las afirmaciones, seguido por piloto del acceso a la plataforma y del funcionamiento del instrumento. También se realizaron los informes de retroalimentación de resultados para cada uno de los expertos, este reporte se mostró las afirmaciones en las que se presentó consenso y las que no, para cada uno de estos dos grupos se presentó las afirmaciones organizadas por categorías conceptuales, con su mínimo, el máximo, la mediana y el RI. En cuanto el instrumento y los reportes estuvieron aprobados se inició con la tercera ronda.

La *tercera ronda* inició con la invitación a los seis expertos a responder el tercer cuestionario, a revisar el informe de retroalimentación y a tenerlo en cuenta al momento de responder el instrumento, junto a este correo electrónico se envió el link de acceso al cuestionario y se adjuntó el reporte individual de resultados en un archivo de formato PDF.

El tiempo que tuvieron los expertos para diligenciar el cuestionario fueron de diez días. Cuatro de los siete expertos contestaron el cuestionario durante este tiempo, respecto a los dos expertos que no respondieron el instrumento, no se pudo establecer contacto con ellos a

pesar de tratar de comunicarse con ellos mediante varios correos electrónicos, por lo que no se conoció la causa de su deserción.

Después de recibir la información de los expertos, se inició la depuración y organización de los datos, seguido de esto se inició con el análisis estadístico, el cual fue el mismo que se realizó en la segunda ronda. Sin embargo, para esta ronda se estimó la estabilidad de las respuestas entre las dos últimas rondas por medio del RIR, estimando, el grado de convergencia de opiniones del grupo (von der Gracht, 2012). También se realizaron los reportes de resultados para cada uno de los expertos participantes en el que se presentó el resumen de los resultados de las dos rondas, para los ítems en los que hubo consenso y en los que no, para cada categoría de análisis.

Resultados

Los análisis de datos se presentan de acuerdo con el orden de las rondas en las que se desarrolló el estudio Delphi, es decir, inicia con los análisis para el cuestionario de pregunta abierta por medio de análisis cualitativos, continúa con los análisis de la segunda ronda para el primer cuestionario de pregunta cerrada y finaliza con los análisis para el último cuestionario cerrado respondido por los expertos, estos últimos con análisis cuantitativos.

Primera ronda

En esta ronda se realizó un análisis de contenido textual de las respuestas de los siete expertos a las seis preguntas abiertas del primer cuestionario. Las categorías resultantes de este análisis se presentan en la tabla 2. Estas categorías y subcategorías son las unidades de análisis de las siguientes rondas.

Tabla 2

Descripción de las categorías teóricas resultantes del análisis de contenido

Categoría	Subcategoría	Definición	Cantidad de ítems
	Concepto de validez	Definición brindada por los expertos sobre el concepto de validez en evaluación en psicología y educación.	8
	Objeto de validez	Opiniones sobre qué es lo que se valida en un proceso de validación.	4
	Consecuencias	Opinión sobre las consecuencias en la definición de validez	4
	Validez en los Estándares (2014)	Opinión sobre la definición de validez dada en los Estándares (2014)	10
	Ideas acerca de los Estándares (2014)	Opiniones generales sobre los estándares 2014, no solo relacionados con el concepto de validez.	4
	Consenso en los Estándares	Opinión sobre el concepto de validez dada en los Estándares (2014) como consenso en la definición de validez.	3
	A favor del consenso	Opiniones que apoyan la idea de que haya consenso en el concepto de validez.	11
	Problemas para alcanzar un consenso	Descripción de algunos de las barreras que se presentan para lograr el consenso.	7

Categoría	Subcategoría	Definición	Cantidad de ítems
	No necesidad de alcanzar un consenso	Opiniones a favor de que no es necesario el consenso en la definición de validez.	5
	Validación	Opiniones sobre el proceso de validación.	9
	Definición de validación	Opiniones sobre el concepto de validación.	9

Segunda ronda

Con el fin de determinar el nivel de acuerdo en cada una de las afirmaciones del segundo cuestionario e identificar si hubo consenso o no entre los seis expertos participantes (Borsboom, Cizek, Kane, Mislevy, Newton y Shaw), se realizaron análisis estadísticos descriptivos para cada una de ellas. En las tablas 3 a 6, se presentan el valor mínimo, el valor máximo, la mediana y el rango intercuartíl (RI) de los ítems que en esta ronda obtuvieron consenso en cada categoría, es decir, de aquellas afirmaciones que tuvieron un $RI \leq 3{,}0$. Los descriptivos de los demás ítems serán tratados en los resultados de la tercera ronda. Vale la pena recordar que un valor del $RI \leq 3{,}0$ indica que al menos la mitad de los expertos expresaron opiniones que no difieren entre sí en más de tres puntos en la escala utilizada; esto no excluye que alguno de los otros expertos haya expresado una opinión por fuera de ese rango, ni que, en ese caso, dicha opinión se encuentre muy apartada de la dada por el grupo dentro del RI.

En la tabla 3 se observa que la afirmación en la que hay mayor acuerdo y consenso es aquella que nombra que las consecuencias de la evaluación con pruebas son una fuente de evidencia para justificar el uso de los puntajes. Respecto al objeto de validez, los expertos están de acuerdo en que es necesario definir qué es la validez para determinar el objeto; sin embargo, la dispersión de las opiniones de esta afirmación es alta. En relación con concepto de la validez, se puede observar que donde hubo mayor acuerdo y consenso, aunque con presencia de posiciones encontradas, es que la validez es el grado en que la evidencia recopilada, la teoría y la argumentación apoyan las inferencias que se pretendan hacer a partir de las puntuaciones de las pruebas, y que estas deben ser significativas, útiles y apropiadas.

Tabla 3

Análisis descriptivos de las afirmaciones en las que hubo consenso de la categoría «concepto de validez», ronda 2

Subcategoría	Afirmación	Descriptivos			
		Mín.	Máx.	Mediana	RI
	Validity is the degree to which it is possible to measure whatever needs to be measured by implementing the assessment procedure for the testing.	2	7	6.0	1.5
	Validity is the extent to which the inferences that are made on the basis of the outcomes of the assessment are meaningful, useful and appropriate.	0	10	7.0	2.3
	Validity is the degree to which collected evidence, theory, and logical argument support the intended inferences to be made from test scores.	0	10	9.5	2.5
Objeto de validez	The proper object of validity cannot be determined unless there is some agreement about what validity actually means.	5	10	8.0	3.0
Consecuencias	The consequences of testing are a source of evidence for justifying the use of a score.	7	10	9.5	1.0

En la tabla 4 se observa que los expertos consideran que la definición de validez que se presenta en los Estándares (2014) es abierta, permite que cualquier interpretación se valide al admitir cualquier tipo de evidencia y propósito para cualquier interpretación, es ambigua porque puede ser interpretada de diferentes maneras y no es clara pues no se puede identificar si es una sola interpretación o varias interpretaciones de los puntajes las que se validan. También consideran que si bien los Estándares (2014) nombran los usos de las puntuaciones como parte de la definición de validez, les hace falta dar más detalles sobre la forma de cómo usar los puntajes, el contexto de su uso y las consecuencias del uso de los puntajes.

Adicionalmente, para los expertos, los Estándares (2014) no son una guía suficiente de validación y consideran que la validez definida como «el grado en que la evidencia y la teoría sustentan» no aplica solo al contexto de la medición en psicología y educación; sino que tiene en un contexto más amplio. Por otro lado, el panel de expertos no está de acuerdo

ni en desacuerdo con la idea de incluir en los Estándares (2014) aspectos relacionados con el desarrollo y los procedimientos de administración de pruebas.

Tabla 4

Análisis descriptivos de las afirmaciones en las que hubo consenso de la categoría «Estándares (2014)», ronda 2

Subcategoría	Afirmación	Descriptivos			
		Mín.	Máx.	Mediana	RI
	Given the definition of validity from the Standards, it is always possible to come up with support for a test score interpretation, especially if one is allowed to introduce any kind of evidence and propose any kind of score interpretation.	0	10	8.0	2.0
	The Standards' definition of validity is limited in the sense that it is ambiguous and can be interpreted in different ways.	1	10	8.0	2.3
	The Standards (2014) focus on interpretations and do not give enough attention to how test scores are used, the contexts of use, and the consequences of use.	4	10	7.5	2.5
	It is unclear in the Standards definition whether the proper object of validity is the intended interpretation of test scores or any number of possible actual interpretations drawn by any number of possible actual users.	1	10	8.0	3.0
	The Standards (2014) are insufficient to guide validation practice.	3	10	7.5	1.8
	Validity, as defined in the Standards in terms of "the degree to which evidence and theory support," is not restricted to educational and psychological measurement, and should be applicable in a wider context.	6	10	8.5	2.5
	Evidence based on "Test Development and Administration Procedures" should be added to the Standards.	5	10	5.5	2.5

En la tabla 5 se observa que, para los expertos, las diferencias entre las aproximaciones respecto al concepto de validez no son pequeñas, y lograr un «consenso en el concepto de validez» ayudaría a mejorar la comunicación entre profesionales y todos los involucrados en la evaluación, así como en las prácticas de validación, al igual que el consenso es

importante porque la definición que se brinde de validez tiene consecuencias en las prácticas de los usuarios en contextos no solo de investigación científica, sino también de evaluación de políticas y asuntos de políticas públicas. Además, el panel de expertos coincide en que existe ambigüedad entre los diferentes actores involucrados en la evaluación en psicología y educación cuando se habla de validez y que el consenso puede ayudar a disminuirla. También están de acuerdo en que para lograr un consenso es importante indagar sobre la razón de porqué los académicos deciden promover diferentes definiciones, más que en el qué se dice sobre validez.

Tabla 5

Análisis descriptivos de las afirmaciones en las que hubo consenso de la categoría «consenso», ronda 2

Subcategoría	Afirmación	Descriptivos			
		Mín.	Máx.	Mediana	RI
Estándares (2014)	There are only small differences among the approaches to defining validity.	0	3	0.5	1.0
	A consensus may be reached by seeking clarity about the reasons why different scholars choose to promote different definitions; that is, not what they are trying to argue about validity but why, which is not always clear.	5	10	7.5	1.0
	It is necessary to know the specific evidence backing the interpretation of test scores as measuring a particular attribute.	7	10	8.0	1.5
	Because test use is intertwined not only with scientific research, but also with policy evaluation and societal political issues, accepting score interpretation or score use as the central question of validity has important consequences for what test users are supposed to do.	5	10	8.5	1.8
	When we do talk about validity –amongst ourselves, to our students, to our stakeholders– there is this slight air of ambiguity over what, exactly, it is that we are supposed to be talking about.	7	10	8.0	2.3
	A consensus definition of validity would help to improve communication among practitioners and stakeholders, as well as to improve validation practice.	1	10	7.0	3.0

Subcategoría	Afirmación	Descriptivos			
		Mín.	Máx.	Mediana	RI
Problemas para lograr consenso	One problem in reaching a consensus over the definition of validity is that the view of policymakers and the public at large remains rooted in historical conceptions that pay less attention to cognitive and social factors.	3	10	5.0	0.8
	One problem in reaching a consensus over the definition of validity is that different stakeholders debate validity from different perspectives, and often want to 'solve' different problems with their preferred definition.	3	10	8.5	1.0
	One problem in reaching a consensus over the definition of validity is that philosophical orientations among both theorists and practitioners are too disparate.	5	10	7.5	1.8
	One problem in reaching a consensus over the definition of validity is that very few measurement specialists have a solid understanding of the educational and psychological measurement literature on validity.	5	10	5.5	1.8
	One problem in reaching a consensus over the definition of validity is that the few scholars who are actively engaged in the topic find themselves drawn into philosophical debates about the concept of validity that then make it harder for non-specialists to access the literature.	1	10	6.0	2.3
	One problem in reaching a consensus over the definition of validity is that scholars approach different kinds of questions (ontological, methodological, ethical).	6	10	8.5	3.0
	We should try to be explicit about how our views of validity follow from antecedent philosophical, methodological, and/or political convictions.	4	9	8.0	1.0
	The differences in approaches to validity cannot be overcome. We have to live with them.	2	9	7.5	2.5

Sin embargo, el panel de expertos está de acuerdo con que existen factores que dificultan alcanzar un consenso, uno de ellos es que las orientaciones filosóficas que guían la definición de validez son muy dispares tanto en la teoría como en lo práctico; con que en este mismo sentido se relaciona la idea que los académicos que estudian la validez se centran en diferentes tipos de preguntas, unos hacen preguntas ontológicas, otros

metodológicas y, otros, éticas, y con que a menudo se quieren solucionar diferentes problemas de la validez desde diferentes perspectivas con su definición preferida.

Por otro lado, los expertos no se mostraron ni de acuerdo ni en desacuerdo con las afirmaciones respecto a que a) la opinión de los políticos y del público en general como enraizada en concepciones del pasado que prestan menos atención a los factores cognitivos y sociales, b) sobre que muy pocos especialistas en medición tienen una sólida comprensión de la literatura de medición educativa y psicológica sobre la validez y, por último, c) sobre que los pocos académicos que participan activamente en el tema se encuentran atraídos en debates filosóficos sobre el concepto de validez que dificulta que los no especialistas accedan a la literatura.

Respecto a la no necesidad de alcanzar un consenso, el panel está muy de acuerdo con que los expertos deben ser explícitos sobre cómo sus opiniones sobre la validez derivan de convicciones filosóficas, metodológicas y / o políticas anteriores y que estas diferencias sobre la validez no pueden ser superadas y por tanto todos tenemos que aprender a vivir con ellas.

Respecto a la validación, en la tabla 6 se puede ver que existe consenso entre el panel de expertos en que están de acuerdo con que la teoría de la validez debe ayudar a comprender el concepto de validez, a qué se refiere, a cómo se relaciona con otros conceptos y a cómo demostrarlo por medio de la validación, también están de acuerdo con que la explicación de las evaluaciones, los contextos de evaluación y las bases de las evidencias debe ser más amplia de lo que es actualmente y en particular, a las extensiones de los procedimientos de validación basados en argumentos, aplicadas a nuevas formas de evaluación, debe proveer ejemplos y establecer las expectativas para el uso adecuado de las evaluaciones.

Tabla 6

Análisis descriptivos de las afirmaciones en las que hubo consenso de la categoría «validación», ronda 2

Subcategoría	Afirmación	Descriptivos			
		Mín.	Máx.	Mediana	RI
	Validity theory should provide a conceptual framework to guide validation practice.	9	10	10.0	0.0
	A theory of validity ought to help us understand the meaning of the concept (validity), what it refers to, how it relates to other concepts and, importantly, how to demonstrate it (validation).	8	10	10.0	0.8
	To improve validation practices, the explication of assessments, contexts, and evidentiary bases should be more comprehensive than it typically is in current practice.	6	10	9.0	2.8
	Extensions of argument-based validation procedures, applied to new forms of assessment, will give examples.	5	10	8.0	2.8
	Extensions of argument-based validation procedures, applied to new forms of assessment, will set expectations for sound use of assessments.	5	10	8.0	2.8
	Validation should state the proposed interpretation and use of scores, and the claims being made, as clearly and completely as is feasible, and then evaluate these claims, preferably by challenging them.	7	10	10.0	0.8
	Validation is achieved when accumulated empirical evidence is enough to show that the inferences based on test scores are reasonable.	0	9	8.5	1.8
	Validation should evaluate the intended test score uses by gathering evidence that supports (or potentially does not support) using the test scores for prespecified purposes.	0	10	8.5	1.8
	Validation consists of gathering evidence that confirms (or potentially disconfirms) an intended score interpretation.	7	10	9.5	2.5
	Validation is achieved when one shows that the test evokes a causal process (e.g., an item response process) such that its outcome indeed depends on the value of the targeted attribute.	3	10	7.0	3.0

También se puede observar en la tabla 6 que hubo un gran consenso y acuerdo entre los expertos con que la validación debe indicar la interpretación propuesta y el uso de los puntajes, y que las afirmaciones que se hacen, deben ser tan claras y completas como sea posible, y luego evaluar estas afirmaciones, preferiblemente desafiándolas, y también con que la validación debe evaluar los usos de la puntuación de la prueba pretendida recopilando evidencia que respalde o no el uso de las calificaciones de las pruebas para propósitos preespecificados. Además, que la validación se logra cuando la evidencia empírica acumulada es suficiente para demostrar que las inferencias basadas en los resultados de las pruebas son razonables, ya que la validación consiste en reunir pruebas que confirman o no la interpretación de la puntuación deseada.

Tercera ronda

Para la última ronda se realizaron los análisis descriptivos para el tercer cuestionario de preguntas cerradas, el cual fue respondido por cuatro expertos (Cizek, Kane, Mislevy, Newton). En la primera parte de estos resultados se presentan los análisis para las afirmaciones por categorías en las que hubo consenso, luego en las que no hubo consenso; y en la segunda parte se realiza la comparación de descriptivos de estas afirmaciones entre la segunda y la tercera ronda para los cuatro expertos que participaron en ambas, y se analiza la estabilidad de las respuestas del grupo.

Afirmaciones en las que hubo consenso. En esta parte se presentan los análisis de acuerdo y consenso para las afirmaciones del tercer cuestionario, organizadas por categorías. En las tablas del 7 al 10 se puede ver que los valores de las puntuaciones de acuerdo en la escala de calificación de este grupo de expertos no son tan altas como los que se vio en la segunda ronda, en la que participaron seis expertos.

Respecto al concepto de validez el panel de expertos está medianamente de acuerdo con que la validez designa la habilidad que tiene un instrumento de medición para detectar la variación entre las entidades medidas en un atributo de interés y con que la validez se refiere al grado en que la evidencia y la teoría respaldan las interpretaciones de las puntuaciones de las pruebas para los usos propuestos de las pruebas.

Tabla 7

Análisis descriptivos de las afirmaciones en las que hubo consenso de la categoría «concepto de validez», ronda 3

Subcategoría	Afirmación	Descriptivos			
		Mín.	Máx.	Mediana	RI
	Validity designates the ability of a measurement instrument to detect variation between measured entities in an attribute of interest.	4	9	6.5	2.0
	Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests.	3	9	7.5	2.3
	The proper object of validity is the interpretation or meaning of a score.	5	8	6.0	0.8
	The object of investigation in validity is the quality of the argumentation and evidentiary backing for a given interpretation and use of scores.	5	10	5.5	2.0
	The proper object of validity is the measurement instrument.	0	6	4.0	3.0
	The consequences of testing could be classified as part of the 'acceptability' of a testing procedure, where validity is nested within acceptability.	7	10	9.0	2.3
	The issue of the defensibility of test score use and the consequences of testing (ethical evaluation) is broader than validity.	5	10	9.0	2.8

En relación con el objeto de validez se observó que los expertos están indecisos frente a la idea de escoger el objeto que se valida, es decir, todavía no es muy claro si son las interpretaciones o significados de los puntajes, o el objeto si es la calidad de la confirmación y el respaldo probatorio para una interpretación y uso de las puntuaciones. También se ve que el panel de expertos están medianamente en desacuerdo con la afirmación de que el objeto de validez es el instrumento de medida.

En referencia a las consecuencias, el panel de expertos alcanza consenso y están muy de acuerdo con que el tema de la defensa del uso de la puntuación de la prueba y las consecuencias de la prueba (evaluación ética) es más amplia que la validez y también con que las consecuencias del uso de las pruebas podrían clasificarse como parte de la "aceptabilidad" de un procedimiento de evaluación, en el que la validez está anidada dentro de la aceptabilidad.

Tabla 8

Análisis descriptivos de las afirmaciones en las que hubo consenso de la categoría «Estándares (2014)», ronda 3

Subcategoría	Afirmación	Descriptivos			
		Mín.	Máx.	Mediana	RI
	The Standards' definition of validity conflates the validity of score interpretation and the appropriateness of score use.	6	10	7.0	1.0
	The importance of validity is widely enough recognized that it finds its way into laws and regulations.	7	10	8.5	1.5
	The Standards for Educational and Psychological Testing (2014) give a proper account of what validity is.	1	7	6.0	1.5
	The Standards (2014) insufficiently recognize the importance of the basic methodological question of whether the test indeed measures the targeted attribute.	2	8	6.5	2.3
Ideas acerca de los Estándares (2014)	The Standards are methodologically weak but politically strong.	0	6	1.0	3.0

En la tabla 8 se ve en relación con la definición dada por los Estándares (2014) de la validez que el panel de expertos están de acuerdo con que la importancia de la validez es tal que incluso orienta leyes y regulaciones. Por otra parte, se observa que están medianamente de acuerdo con que la definición de validez brindada por los Estándares (2014) confunde la interpretación de los puntajes con la adecuación del uso de las puntuaciones, con que no reconocen la importancia de la pregunta metodológica básica sobre si la prueba mide realmente el atributo objetivo, y por último con que los Estándares (2014) dan cuenta apropiadamente de lo que es la validez.

Por último, el panel de expertos está en desacuerdo con la afirmación de que los Estándares (2014) son metodológicamente débiles, pero políticamente fuertes.

Tabla 9

Análisis descriptivos de las afirmaciones en las que hubo consenso de la categoría «consenso», ronda 3

Subcategoría	Afirmación	Descriptivos			
		Mín.	Máx.	Mediana	RI
	The Standards (2014) achieved consensus on defining validity.	0	5	2.0	2.8
	The Standards' definition of validity is limited in the sense that respected scholars (and who-knows-how-many measurement specialists and practitioners) do not fully accept this definition.	4	10	9.0	3.0
	It is necessary to know whether the intended practical use of the test scores is justified.	8	10	9.0	0.5
	A way to reach agreement on some definitional issues on validity would be to implement discussion strategies focusing on specific points.	5	6	5.5	1.0
	It is necessary to know what a test measures.	5	10	8.5	2.0
	Advances in technology make it possible to carry out highly innovative forms of assessment. To not have a conception of validity that applies to these forms of assessment is to invite misinterpretations and unsound practices.	2	10	7.0	2.0
	If there is no consensus about the meaning of validity (whether by formal definition or by the way it is used), then effective communication is not possible.	3	9	6.5	2.3
	Differences in views about validity primarily represent differences of opinion about how the field of educational measurement should function now and in the future.	0	7	6.5	2.5
Problemas para alcanzar un consenso	One problem in reaching a consensus over the definition of validity is that there are too many theorists with perspectives who don't have practical experience doing validation work.	3	9	4.5	2.3
	The differences in approaches to validity should not be overcome. We should embrace a certain amount of pluralism.	2	8	4.5	2.3
	The differences in approaches to validity should not be overcome. The debates about validity arise because it is a foundational concept.	2	8	5.0	3.0

En la tabla 9 se observa que, respecto al consenso del concepto de validez en los Estándares (2014), los expertos están de acuerdo en que la definición de validez provista por los Estándares es limitada en cuanto a que los académicos, especialistas en medición y demás usuarios no aceptan totalmente esta definición, y están en desacuerdo con la afirmación que los Estándares (2014) alcanzan un consenso sobre el concepto de validez.

En relación con las razones por las que sería interesante lograr un «consenso sobre el concepto de validez», se logró consenso en que están de acuerdo con que es necesario saber si el uso práctico de los resultados de las pruebas está justificado y saber lo que mide una prueba. También llegaron al consenso de que están medianamente de acuerdo con que los avances tecnológicos permiten llevar a cabo formas de evaluación altamente innovadoras y al no tener una concepción de validez que se aplique a estas formas de evaluación se invita a hacer interpretaciones y prácticas erróneas, con que si no hay consenso sobre el significado de la validez (ya sea por definición formal o por la forma en que se utiliza), entonces la comunicación efectiva no es posible y, por último, con que las diferencias en las opiniones acerca de la validez representan principalmente diferencias de opinión sobre cómo el campo de la medición educativa debe funcionar ahora y en el futuro. Finalmente, en esta subcategoría, el panel de expertos están indecisos con la idea que una forma de llegar a un acuerdo sobre algunas cuestiones de definición sobre la validez sería implementar estrategias de discusión enfocadas a abordar puntos específicos.

Sobre los problemas para alcanzar un consenso, los expertos tiene una posición neutra respecto a la idea de que hay demasiados teóricos con perspectivas que no tienen experiencia práctica haciendo trabajo de validación.

Respecto a las ideas sobre que no hay necesidad de consenso, los expertos están indecisos con la afirmación sobre que las diferencias en los enfoques de validez no deben ser superadas, sino que debemos adoptar una cierta cantidad de pluralismo y con que estos debates sobre la validez surgen porque este es un concepto fundamental.

Tabla 10

Análisis descriptivos de las afirmaciones en las que hubo consenso de la categoría «validación», ronda 3

Subcategoría	Afirmación	Descriptivos			
		Mín.	Máx.	Mediana	RI
	If validity theory is confused or contested, then emerging frameworks will lack clarity in the practical guidance required for their execution.	6	8	7.5	1.3
	One might adopt a fairly narrow definition of 'validity' while still adopting a far broader definition of 'validation'.	4	9	7.0	2.8
	Validation would include all sorts of sources of empirical evidence and logical analysis, including evidence of the consequences of testing.	0	10	9.0	2.5
	The social value of an assessment should be evaluated, but not as part of its validation.	5	10	9.0	1.3

En la tabla 10 se observa que el grupo de expertos están medianamente de acuerdo con que, si la teoría de la validez se confunde o se discute, entonces los marcos emergentes carecerán de claridad en la guía práctica requerida para su ejecución y con que podría adoptarse una definición bastante estrecha de «validez», al tiempo que adopta una definición mucho más amplia de «validación».

También se observa que, si bien alcanzaron consenso en estar de acuerdo con que el valor social de una evaluación debe ser evaluado, pero no como parte de su validación; también llegaron a estar de acuerdo con que la validación incluiría todo tipo de fuentes de evidencia empírica y análisis lógico, incluyendo pruebas de las consecuencias de la medida.

Afirmaciones en las que NO hubo consenso. En esta parte se presentan los análisis de la tercera ronda para las afirmaciones que no alcanzaron consenso en la segunda ni tercera ronda, organizadas por categorías. Es decir, que respecto a estos temas los expertos tienen opiniones muy diferentes respecto a cada una de ellas.

Tabla 11

Análisis descriptivos de las afirmaciones en las que NO hubo consenso de la categoría «concepto de validez», ronda 3

Subcategoría	Afirmación	Descriptivos			
		Mín.	Máx.	Mediana	RI
	Validity is a socially situated process that encompasses the uses and consequences of the measurements evaluated.	0	10	7.5	3.3
	Validity is the extent to which the claims (interpretations and score-based decisions) based on assessment scores are adequately supported by evidence.	3	10	8.5	4.0
	Validity is not intended to encompass ethical evaluation of test score use.	4	10	7.0	4.5
Consecuencias	The consequences of testing are not a source of validity evidence.	1	10	4.0	3.8

En la tabla 11 se muestra que respecto al concepto de validez, los expertos tienen posturas muy diferentes sobre las ideas que la validez no pretende abarcar la evaluación ética del uso de la puntuación de la prueba, a que la validez es la medida en que las afirmaciones (interpretaciones y decisiones basadas en la puntuación) basadas en las puntuaciones de la evaluación están adecuadamente apoyadas por la evidencia, a que la validez es un proceso socialmente situado que abarca los usos y consecuencias de las mediciones evaluadas y a que las consecuencias de las pruebas no son una fuente de evidencia de validez.

Tabla 12

Análisis descriptivos de las afirmaciones en las que NO hubo consenso de la categoría de «Estándares (2014)», ronda 3

Subcategoría	Afirmación	Descriptivos			
		Mín.	Máx.	Mediana	RI
	The consequences of testing should be removed from the Standards' definition of validity.	0	10	4.5	3.3
	The continuing debate about the concept of validity is less important than a shared understanding of validation.	1	9	6.0	3.5

En la tabla 12 se ve que respecto a los Estándares (2014) el panel de expertos no llegó a un consenso sobre las consecuencias de las pruebas deben eliminarse de la definición de validez de los Estándares y con que el debate continuo sobre el concepto de validez es menos importante que una comprensión compartida de la validación.

En la tabla 13 se observa respecto al «consenso en el concepto de validez», los expertos tienen muy diferentes opiniones frente a la idea que sobre que cualquier definición particular de validez afectará la definición de todo tipo de otros conceptos de «calidad de evaluación» (por ejemplo, confiabilidad, imparcialidad). Por lo tanto, debemos pensar en términos de definir conjuntos de «conceptos de calidad» dentro de «marcos de calidad» en lugar de pensar simplemente en términos de definir un solo «concepto de calidad», la validez.

Tabla 13

Análisis descriptivos de las afirmaciones en las que NO hubo consenso de la categoría «consenso», ronda 3

Subcategoría	Afirmación	Descriptivos			
		Mín.	Máx.	Mediana	RI
No necesidad de un consenso	Any particular definition of validity will affect the definition of all sorts of other 'assessment quality' concepts (e.g. reliability, fairness). So, we should think in terms of defining sets of 'quality concepts' within 'quality frameworks' rather than thinking simply in terms of defining a single 'quality concept', validity.	2	10	8.0	3.5

En relación con la validación, la tabla 14 se muestra que los expertos no llegaron a un consenso sobre las ideas que es muy difícil juzgar qué tipo de evidencia de validación es necesaria y que es muy difícil juzgar cuánta evidencia de validación es suficiente.

Tampoco se alcanzó un consenso respecto a que la validación debe abarcar cualquier cosa relacionada con la evaluación científica de la «calidad de la medición» para un procedimiento de evaluación particular y a que, durante la validación, las consecuencias de las pruebas deben ser evaluadas sólo con respecto a cómo afectan el significado de las puntuaciones.

Tabla 14

Análisis descriptivos de las afirmaciones en las que NO hubo consenso de la categoría «validación», ronda 3

Subcategoría	Afirmación	Descriptivos			
		Mín.	Máx.	Mediana	RI
	It is very hard to judge what kind of validation evidence is necessary.	2	10	4.5	4.3
	It is very hard to judge how much validation evidence is sufficient.	2	10	5.0	5.0
	Validation should embrace anything related to the scientific evaluation of 'measurement quality' for a particular assessment procedure.	3	10	6.0	3.3
	During validation, the consequences of testing should be evaluated only with respect to how they affect the meaning of scores.	4	10	7.5	5.3

Comparación entre segunda y tercera ronda. En esta parte de los resultados se presenta el análisis de la estabilidad de las respuestas del grupo, comparando el RIR con la diferencia de medianas de la tercera ronda respecto a la segunda ronda. Para estos análisis solo se tuvieron en cuenta las respuestas dadas por los cuatro expertos que participaron en la tercera ronda, se compararon sus respuestas a las 38 afirmaciones que no tuvieron consenso en la segunda ronda y que se preguntaron nuevamente en la tercera ronda. Cada una de estas afirmaciones es representada por un punto en la figura 1, en la cual se identifica la categoría a la que pertenece la afirmación y si los cuatro expertos mostraron un consenso en la tercera ronda.

En la figura 1, el eje X muestra la diferencia de las medianas entre la tercera y segunda ronda; cuanto más alejado del 0 mayor cambio en el nivel de acuerdo por parte de los cuatro participantes entre las rondas segunda y tercera en la afirmación. El eje Y muestra el cambio relativo en el rango intercuartíl de las afirmaciones entre la tercera y segunda rondas. Cuanto más lejos del 0 esté el punto que representa cada afirmación, mayor cambio respecto al consenso.

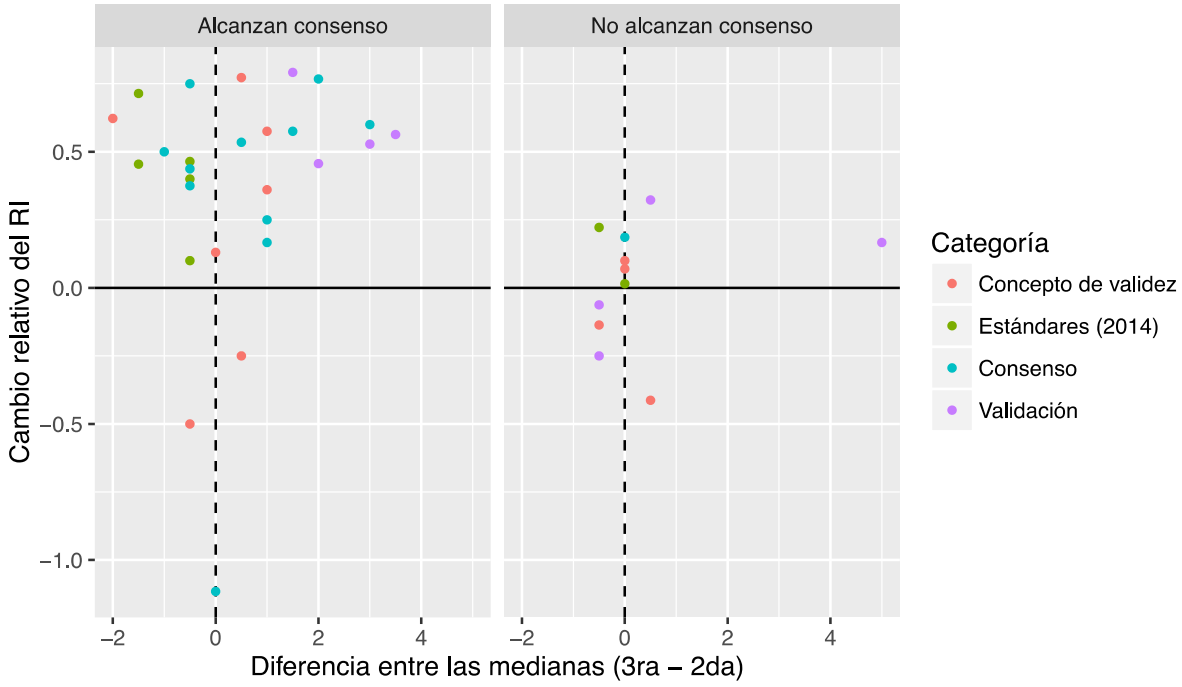


Figura 2. Estabilidad de las respuestas del grupo entre la segunda y tercera ronda contra el cambio del nivel de acuerdo para estas mismas rondas.

En la figura 1 se puede observar que las afirmaciones sobre el concepto de validez presentaron niveles de acuerdo similares en ambas rondas y que las afirmaciones sobre los Estándares (2014) tuvieron menor nivel de acuerdo en la tercera ronda. Por su parte, para las afirmaciones sobre el consenso y sobre los Estándares (2014) se tuvo un cambio más consistente entre los expertos, quienes llegaron a un mayor consenso en la tercera ronda; sin embargo, vale la pena anotar que para algunas afirmaciones los cuatro expertos ya habían logrado un consenso en la segunda ronda, aunque este consenso no se encontraba entre los seis expertos que participaron en ella, y que aun cuando sus respuestas fueron mucho menos homogéneas en la tercera ronda (RIR = -1.2), el consenso entre ellos se mantuvo. Adicionalmente, se puede apreciar que para las afirmaciones sobre validación en las cuales se llegó a un consenso en la tercera ronda, los expertos se mostraron no sólo más homogéneos (lo cual permitió que llegaran a un consenso sobre ellas), sino que además se mostraron más de acuerdo con estas afirmaciones.

Conclusiones

En este proyecto se lograron identificar, entre un grupo de expertos, los puntos en los que existe consenso y aquellos en los que no sobre el concepto de validez en educación y psicología, por medio el desarrollo de un estudio Delphi en línea de tres rondas. En este apartado se presentan las conclusiones derivadas del desarrollo de este estudio. En primer lugar, se presenta la descripción de las categorías de análisis derivadas de la primera ronda que estructuran el contenido del análisis de los resultados durante todo el estudio; después se muestra el análisis general de los resultados del estudio Delphi, con énfasis en los puntos en lo que no hay consenso, seguido del análisis de las ideas más relevantes en las que se alcanza consenso por parte del panel de expertos para cada categoría de análisis. Por último, se describen las implicaciones y las recomendaciones que sugiere este trabajo con el fin de aportar elementos que contribuyan a dar claridad y consistencia a la conceptualización de la validez.

Las primeras actividades que se realizaron fueron identificar las principales preguntas que se han abordado en el desarrollo de la teoría de la validez en las últimas décadas y las posturas de expertos en cada uno de ellos, e identificar las categorías de análisis alrededor de las cuales se orientó la discusión y la elaboración de los cuestionarios con los que se buscó encontrar consenso entre los expertos participantes. Las categorías de análisis definidas a partir de la revisión bibliográfica y de las opiniones expresadas por los expertos durante la primera ronda del estudio fueron:

1. Concepto de validez. En esta categoría se describe la definición brindada por los expertos sobre el concepto de validez en evaluación en psicología y educación, las opiniones sobre qué es lo que se valida en un proceso de validación, y la opinión sobre las consecuencias en la definición de validez.
2. Estándares (2014). En esta categoría se presenta la opinión sobre la definición de validez dada en los Estándares (2014), así como las opiniones que los expertos tienen sobre los Estándares (2014) en general, no sólo relacionados con el concepto de validez.
3. Consenso. En esta categoría se exponen las opiniones del panel de expertos sobre el concepto de validez dada en los Estándares (2014) como consenso en la definición de validez, las opiniones de quienes apoyan la idea de que haya consenso en el concepto de

validez, la descripción de algunos de las barreras que se presentan para lograr el consenso y las opiniones a favor de que no es necesario el consenso en la definición de validez.

4. Validación. En esta categoría se describen las opiniones de los expertos sobre el proceso de validación y sobre el concepto de validación.

Adicionalmente, por medio de un estudio Delphi se identificaron los elementos sobre el concepto de validez, en los que los expertos participantes llegaron a un consenso y aquellos en los que no. Estos elementos están estructurados según las categorías de análisis, obtenidas de las opiniones expresadas por ellos mismos durante la primera ronda del estudio. En la tabla 15 se presenta la cantidad de afirmaciones en las que se alcanzó consenso entre los expertos, tanto para la segunda ronda como para la tercera y la cantidad de afirmaciones en las que no se logró un consenso.

Tabla 15

Descripción de resultados generales del consenso en las rondas del estudio Delphi

Categoría	Subcategoría	Cantidad	Consenso ronda 2	Consenso ronda 3	No Consenso
	Concepto de validez	8	3	2	3
	Objeto de validez	4	1	3	
	Consecuencias	4	1	2	1
	Validez en los Estándares	10	4	4	2
	Ideas acerca los Estándares (2014)	4	3	1	
	Consenso en los Estándares (2014)	3	1	2	
	Proconsenso	11	5	6	
	Problemas para lograr el consenso	7	6	1	
	No necesidad del consenso	5	2	2	1
	Validación	9	5	2	2
	Definición de validación	9	5	2	2
Total		74	36	27	11

La tabla 15 muestra que, al finalizar la tercera ronda del estudio, se alcanzó consenso en 63 de 74 afirmaciones relacionadas con la validez. La subcategoría de análisis en la que hubo un menor consenso fue «concepto de validez», ya que de ocho afirmaciones solo hay consenso en cinco. Por otro lado, en la categoría que se logró mayor consenso es «consenso»; de 26 afirmaciones, solo en una no hay acuerdo entre los expertos.

A lo largo del estudio, la opinión sobre las afirmaciones acerca de un «consenso en el concepto de validez» se mostró más estable que las opiniones sobre los Estándares (2014), esto refleja que para la categoría de «Estándares (2014)» ya había consenso entre los expertos con anterioridad al desarrollo del estudio Delphi mientras que en la categoría de «validación» se llegó al consenso mediante el desarrollo del estudio Delphi.

Estos resultados generales muestran que en la categoría en «concepto de validez», los expertos no alcanzan un consenso respecto al papel de las consecuencias dentro de la validez, existen opiniones muy variadas sobre la idea de que la validez es un proceso socialmente situado que abarca tanto los usos como las consecuencias de las medidas utilizadas, que la validez no pretende incluir la evaluación ética del uso de los test y que las consecuencias del uso de los test no son una fuente de evidencia de validez. Estos aspectos son algunos de los puntos muertos en la discusión sobre el concepto de validez. Las opiniones sobre si las consecuencias del uso de la prueba deben ser eliminadas de los Estándares o deben ser evaluadas solo respecto a cómo estas afectan el significado de los puntajes de la prueba tampoco logran un consenso de los expertos. Estos pueden ser esos puntos muertos, en los que no hay acuerdo ni consenso general, que nombran Newton y Shaw (2016). En este sentido, el rol de las consecuencias en la validez no es claro todavía; sin embargo, existe una idea respecto a las consecuencias en la que los expertos llegaron a estar de acuerdo en conjunto y es que las consecuencias del uso de las pruebas son una fuente de evidencia para justificar el uso del puntaje. Este puede ser un punto de partida para abrir la discusión sobre cómo tratar las consecuencias del uso de las pruebas, destrabar los puntos muertos relacionadas con estas y fortalecer metodológica y teóricamente el desarrollo del estudio y la práctica del análisis de las consecuencias, que sigue siendo un aspecto de vital importancia para la teoría validez, tal como lo sugiere Cizek (2010b, 2012, 2016).

Otros aspectos en los que se presentaron opiniones muy dispares, después de dos rondas de preguntas, fueron si la continuación del debate sobre el concepto de validez era menos importante que entender la validación y si se debía pensar en términos de definir conceptos y estructuras, en lugar de pensar en términos para definir un único concepto de calidad como lo es la *validez*. Estas ideas no son muy claras o aceptadas de forma homogénea por parte del grupo de expertos, es posible que se quiera continuar con el debate sobre la validez y estas nuevas propuestas de ver su conceptualización.

Otra de las ideas en la que se presentó una gama amplia de opiniones, sin llegar a un consenso, fue sobre qué tan difícil es identificar cuánta evidencia es necesaria y suficiente para validar una prueba. A pesar de la gran cantidad de experiencia en diversas áreas que tienen los expertos, para algunos es sencilla esta tarea porque es posible que sigan protocolos establecidos después de muchos años de trabajo en el área, mientras que para otros no es tan clara su resolución si se tiene en cuenta a todos los usuarios, psicómetras, comercializadores y demás personas involucradas en el área.

Finalmente, hubo dos ideas muy específicas en las que no hubo consenso, la primera se refiere a la definición de validez, dada desde la propuesta de que la validez es la medida en que las afirmaciones (interpretaciones y decisiones basadas en la puntuación) basadas en las puntuaciones de la evaluación están adecuadamente apoyadas por la evidencia; y la segunda, que la validación debe abarcar cualquier cosa relacionada con la evaluación científica de la «calidad de la medición» para cada procedimiento de evaluación en particular. Estas dos ideas cubren una amplia gama de opiniones entre los expertos, lo que puede interpretarse como que estas no son lo suficientemente precisas o suficientes para definir la validez y la validación.

Por otra parte, sobre las ideas en las que se llegó a un consenso es importante anotar que, en general se observa, entre los expertos en este estudio que, si bien se alcanzó una tendencia a lograr una opinión homogénea de una mayoría, también fue usual que se presentara la opinión de uno o dos expertos que no están de acuerdo con esta mayoría. Esto indica que las ideas que se presentan a continuación no son una generalización de consenso entre todos los expertos participantes, sino que pueden ser vistas como ideas en los que no difieren demasiado en sus opiniones y pueden ser fuente para lograr un trabajo en común

para robustecer la teoría de la validez, que den mayor claridad en el uso de conceptos para comunicarse de manera más eficiente con todos los profesionales involucrados en los procesos de evaluación, ayuden a generar un diálogo más concreto sobre las ideas en las que hay mayor controversia y quizás permitan llegar a acuerdos más robustos en estos puntos de debate.

La definición sobre el concepto de validez en la que se presenta mayor acuerdo entre los expertos es «*la validez es el grado en el cual la evidencia recolectada, la teoría y la argumentación apoyan las inferencias que se pretenden hacer a partir de las puntuaciones*». A esta le sigue que la «*validez se refiere a qué tanto las inferencias que son derivadas de los resultados de una evaluación sean significativas, útiles y apropiadas*». Estas definiciones, en las que se encontró consenso, se centran en las inferencias de los puntajes y retoman casi textualmente las definiciones dadas por los Estándares de los años 2014, 1999 y 1985 cuando introducen el concepto de validez. Es decir que, si bien para los expertos la definición de *validez* dada por los Estándares (2014) es ambigua y no logra un consenso, estos constituyen la principal referencia para la conceptualización de la *validez*. También se puede afirmar que a pesar de que hubo consenso en estar de acuerdo con estas definiciones de validez, todavía se considera inexacta o incompleta de alguna manera pues no es unánime el consenso. Por otra parte, a las definiciones que apuntan hacia que la validez se refiere a la habilidad del instrumento de detectar la variación en las entidades de medida en un atributo o que la validez es el grado en que la evidencia y la teoría respaldan las interpretaciones de los puntajes de las pruebas para los usos propuestos de la prueba no logró un nivel de acuerdo ni de consenso entre los expertos muy alto. Esto podría indicar, por una parte, que considerar la validez centrada en el instrumento y su relación causal con el atributo no logra ser una definición que abarque toda la gama de posibilidades de medición cuando se trata de atributos de estudio en la psicología y la educación, tal como lo señala Kane (2016); y respecto a la segunda definición, la única diferencia con las que tuvieron un mayor consenso y acuerdo en la segunda ronda, es que en esta se nombran los usos propuestos de la prueba, esto deja ver que los expertos llegan a un acuerdo más sólido respecto al rol de las inferencias en la conceptualización de la validez, mas no sucede lo mismo con el uso de estos puntajes.

Respecto al objeto de la validez, los expertos están de acuerdo en que es difícil establecer cuál es, ya que depende mucho de una definición más clara de *validez*; y no logran escoger como objeto de la validez entre la interpretación de los puntajes, la calidad de la argumentación y evidencia que subyace a la interpretación y uso de los puntajes, y el instrumento de medida.

En relación con el papel de las consecuencias en la definición de *validez*, se observó que hay un gran consenso y acuerdo con que las consecuencias son una fuente para justificar los usos de las pruebas, y que el tema de la defensa del uso de los puntajes y las consecuencias de estos, visto como una evaluación ética, es más amplio que el tema de la *validez*. Este resultado consolida la posibilidad de considerar la postura de Cizek (2010a, 2010b, 2012, 2016a) respecto a tratar las consecuencias del uso de las pruebas aparte de la validez, y empezar a trabajar en su conceptualización y robustecimiento metodológico desde la justificación.

Respecto a los Estándares (2014), a pesar de cubrir una amplia gama de opiniones, se llega a establecerse una opinión mayoritaria entre los expertos que están de acuerdo con que la definición de *validez* dada en los mismos es ambigua, que permite tantas interpretaciones como usuarios lo que esto permite validar cualquier instrumento y que es limitada porque no todos los académicos ni todos los profesionales y usuarios aceptan esta definición. Esto sugiere que se haga una revisión de esta definición para que sea más clara para todos los involucrados en los procesos de evaluación, diseño y validación de pruebas. Algunos aspectos que pueden ayudar a dar claridad a la definición son a) disolver la mezcla en la definición que hay sobre la validez de las interpretaciones y la adecuación del uso de las pruebas; y b) brindar coherencia conceptual y metodológica entre la definición de *validez* y las fuentes de validez, pues los Estándares (2014) se centran más en el estudio de las interpretaciones del puntaje que en el uso, es decir, el estudio de los contextos y las consecuencias del uso de las pruebas es pobre, a pesar que los usos son parte de la definición de validez brindada en estos.

En relación con la idea de llegar a un «consenso en la definición de la validez», los expertos estuvieron de acuerdo entre ellos respecto a que las diferencias sobre la definición

de la validez no son pequeñas, y que estas diferencias no necesariamente pueden ser superadas y que se tendrá que vivir con estas diferencias. Esto podría apoyar la opción de aceptar la ambigüedad del concepto de validez, sugerido por Newton y Shaw (2016). Sin embargo, hubo un alto acuerdo entre los expertos sobre que el *consenso* en el concepto de validez ayuda a mejorar la comunicación entre los usuarios y todos los involucrados en los procesos de evaluación, a disminuir la ambigüedad conceptual entre académicos, profesionales y usuarios, y a optimizar las prácticas de validación. Ellos también están de acuerdo con la idea que para lograr un *consenso* es vital saber la razón por la cual los académicos escogen promover diferentes definiciones de *validez* y lograr que los académicos y demás profesionales sean explícitos en sus antecedentes filosóficos, metodológicos y convicciones políticas que influyen en sus definiciones; ya que tanto los usuarios como los académicos tienen aproximaciones diferentes para ver la validez (ontológicas, metodológicas o éticas), y esto promueve que se hallen soluciones a los problemas de validación basados en su perspectiva favorita. Esto sugiere que los expertos reconocen la importancia de encontrar un *consenso* entre los académicos que estudian la teoría de validez, y están de acuerdo con trabajar a favor de alcanzar un *consenso* sobre la definición de validez, no solo entre los académicos sino también con todos los profesionales involucrados en los procesos de evaluación y de validación de pruebas. Es decir, la tendencia es hacia que se trabaje por un *consenso* en la definición del concepto de validez, de tal manera que es posible que las probabilidades de lograr un consenso no sean tan bajas como lo proponen Newton y Shaw (2016); sin embargo, en la negociación para lograr un consenso sí es necesario ceder en ciertas ideas que en la actualidad no tienen tanta aceptación y acoger otras nuevas para contribuir en el desarrollo de la teoría y las practicas involucradas en los procesos de medición y evaluación tanto en la educación como en la psicología.

Aunque hubo ideas en las que los expertos coincidieron en que no saben si son ciertas o no, estos pueden considerarse aspectos a explorar con mayor detalle, ideas tales como que los profesionales que trabajan en políticas públicas mantienen el concepto de validez dado en el pasado que presta menos atención a los factores cognitivos y sociales involucrados en la evaluación, o que existen muy pocos especialistas en medición que tenga un sólido

entendimiento de la literatura de validez en la medición en psicología y educación o que alguno académicos que están estudian en la teoría de validez se concentran en debates filosóficos que hacen muy difícil el acceso a la literatura sobre validez por parte de los profesionales que no son especialistas. Saber qué tanto estas apreciaciones son ciertas permite que se encuentren aspectos concretos en los que se puedan trabajar a favor del consenso entre todos los interesados en la teoría de validez.

Los expertos también nombraron aspectos teóricos en los que estuvieron de acuerdo entre ellos y que facilitan el «consenso en el concepto de validez». Estas ideas en las que coincidieron en estar de acuerdo son a) que es necesario conocer tanto la evidencia que sustenta una interpretación como la evidencia que respalda la medida del atributo, b) que la pregunta sobre el concepto de validez es muy importante porque si se opta ya sea por la interpretación de los puntajes o por los usos de los puntajes, esta decisión tiene implicaciones no solo en las investigaciones en el campo, sino en la evaluación de políticas y en políticas públicas en general, ya que estas orientan lo que se supone deben hacer los usuarios, y c) que es necesario conocer si el uso práctico propuesto de la prueba está justificado. Estos tres elementos han sido objeto de debate en diferentes publicaciones entre los expertos y queda claro que para ellos todos estos son aspectos que se deben tener en cuenta en la teoría de la validez, el problema radica en el peso que cada experto le da a cada uno; por lo tanto, es importante reconocer que entre todos no hay acuerdos básicos y que es necesario negociar el peso de estos a favor de lograr mayor claridad y consistencia en la definición.

En la categoría de «validación» los expertos dieron respuestas más homogéneas que en las demás categorías de análisis. Sobre esta categoría se puede afirmar que el panel de expertos está más de acuerdo con realizar una estructura que guíe el proceso de validación y que la teoría de la validez debiera sustentar esta estructura, que en la búsqueda del mejoramiento del proceso de validación se recomienda que estén más conectadas la recolección de evidencia empírica con el análisis lógico tanto para las interpretaciones del puntaje como para la justificación de las consecuencias del uso de la prueba y que la explicación de las evaluaciones, contextos y las evidencias que respaldan la validez de la

prueba sean más completas y exhaustivas que lo que usualmente son. Sin embargo, es notorio que hubo al menos un experto que opinó que la validación de una prueba no se garantiza con evidencia de que las inferencias son razonables, ni evaluando los usos ni las consecuencias de los usos de la prueba, ni es suficiente solo cuando se tiene una gran cantidad de evidencia empírica.

Por otro lado, aunque la mayoría de los expertos están de acuerdo con tomar como referencia la aproximación basada en argumentos para dirigir el proceso de validación y solicitan más ejemplos y la definición de sus alcances a favor de mejorar la validación, entre los expertos no hubo consenso sobre la definición de validez que ofrece esta aproximación. Este es un elemento que crea disonancia, porque todos los expertos participantes están de acuerdo en que la teoría de validez debe proveer la estructura conceptual para guiar la práctica de validación, y si bien están de acuerdo con la guía para la validación propuesta por aproximación basada en argumentos, no están de acuerdo con su conceptualización.

Finalmente, se puede nombrar como limitaciones de este trabajo que, aunque se invitaron a representantes de las posturas que han liderado la discusión, no se pudo contar con la participación de todos ellos. Esto puede limitar el alcance de las conclusiones derivadas de este trabajo; sin embargo, los puntos más importantes de la discusión del concepto de validez fueron tratados por las diferentes posturas fueron abordados en el estudio.

En general se puede afirmar que este trabajo respondió a la discusión actual del concepto de validez siguiendo las conclusiones de las últimas publicaciones y presentaciones de expertos en teoría de validez. También que este trabajo aportó, desde la metodología Delphi, otra estrategia de discusión, que logró medir de alguna forma que tanto están en desacuerdo o de acuerdo con ideas que no han sido trabajados a profundidad por todos por igual. Y específicamente, respecto al concepto de validez se logró identificar las posturas claramente argumentadas respecto al concepto de validez y los puntos en los que hay consenso y en los que no para diferentes aspectos de la conceptualización de la validez y también se logró encontrar recomendaciones para cada una de las categorías, las

cuales permiten nutrir a futuro tanto teórica como metodológicamente la validez y los procesos de validación.

Referencias

Aichholzer, G. (2009). The Delphi Method: Eliciting Expert's Knowlodge in Technology Foresignt. In: A. Bogner, B. Liting y W. Menz (Ed.) *Interviewing Experts* (pp. 252-274). England. Palgrave Macmillan.

Allen, M. J. y Yen, W. M. (2002). *Introduction to measurement theory*. Long grove, Illinois, USA: Waveland Press Inc. (Original work published in 1979)

American Educational Research Association. (1955). Technical Recommedations for achievement tests. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC, USA: American Psychological Association.

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC, USA: American Psychological Association.

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC, USA: American Psychological Association.

American Psychological Association, American Educational Research Association y National Council on Measurement in Education. (1974). *Standards for educational and psychological test.* Washington, DC, USA: American Psychological Association.

Anastasi, A. (1938). Faculties versus factors: A reply to professor Thurstone. *Psychological Bulletin, 35,* 391-395.

Arias, E. M. (2008). *Detección de DIF con Estadísticos basados en Tablas de Contingencia: El Mantel-Haenszel* (Tesis de maestría). Universidad Nacional de Colombia, Bogotá.

Barajas, R. (2017). *Validez en Test Adaptativos Informatizados: Evidencia en un TAI diseñado para evaluar compresión lectora en personas con y sin limitación visual* (Tesis de maestría). Universidad Nacional de Colombia, Bogotá.

Bechtoldt, H. P. (1951). Selection. En S. S. Stevens (Ed.). *Handbook of Experimental Psychology*. (pp. 1237-1267). Nueva York: Wiley.

Berrío, A. I. (2008). *Efecto de la Razón de Tamaños y Desajustes al Modelo en la Detección de Ítems con Funcionamiento Diferencial mediante Procedimientos basados en IRT (Diferencia de dificultad y χ^2 de Lord)* (Tesis de maestría). Universidad Nacional de Colombia, Bogotá.

Binet, A. (1905). New methods for the diagnosis of the intelectual level of subnormals. *L'Annnée Psychologique, 12*, 191-244.

Binet, A. y Henri, B. (1899). La psychologic individualle. *Amiee Psychol., 2,* 2411-465.

Birko, S., Dove, E. S., y Özdemir, V. (2015). Evaluation of Nine Consensus Indices in Delphi Foresight Research and Their Dependency on Delphi Survey Characteristics: A Simulation Study and Debate on Delphi Design and Interpretation. *PLoS ONE*, 10(8) e0135162. Recuperado el 10 de marzo, 2017 de http://doi.org/10.1371/journal.pone.0135162

Bonnemaizon, A., Cova, B. y Louyot, M-C. (2007). Relationship Marketing in 2015: A Delphi Approach. *European Management Journal, 25,*50-59.

Borsboom, D., Mellenbergh, G. J. y Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061-1071.

Borsboom. D., Cramer. A.O.J., Kievit, R. A., Scholten, A.Z. y Franic, S. (2009). The end of construct of validity. En: R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications.* (pp. 135-170). United States of America: Information Age Publishing, INC.

Borsboom, D. (2012). Whose consensus is it anyway? Scientific versus legalistic conceptionsof validity. *Measurement: Interdisciplinary Research and Perspectives, 10,* 38-41.

Borsboom, D. y Wijsen. L. (2016). Frankenstein's validity Mosnter: the value of keeping politics and science separated. A*ssessment in Education: Principles, Policy and Practice, 23,* 281-283.

Bravo, M. de L. y Arrieta, J. J. (2005). El método Delphi. Su implementación en una estrategia didáctica para la enseñanza de las demostraciones geométricas. *Revista Iberoamericana de Educación*, 35 (3). Recuperado el 22 de junio, 2016 de www.rieoei.org/inv_edu38.html.

Brown, F. (1980). *Principios de la medición en psicología y educación.* México: Manual Moderno.

Buck, A. J., Gross, M., Hakim, S. y Weinblatt, J. (1993) Using the Delphi process to analyse social policy implementation – a post hoc case from vocational rehabilitation. *Policy Sciences*, *26* (4), 271–288.

Buckingham, B. R., McCall, W. A., Otis, A. S., Rugg, H. O., Trabue, M. R. y Courtis, S. A. (1921). Report of the Standardization Committee. *Journal of Educational Research, 4(1),* 78-80.

Bunge, M. (1973). On confusing 'Measure' with 'Measurement' in the methodology of behavioral science. En M. Bunge (Ed.), *The methodological unity of science.* (pp 105-122). Dordrecht, The Netherlands: D. Reidel Publishing Company.

Bunge, M. (1985). *La investigación científica.* Barcelona: Ariel S. A.

Campos, V., Melían, A. y Sanchis J. R. (2014). El método Delphi como técnica de diagnóstico estratégico. Estudio empírico aplicado a las empresas de inserción en España. *Revista Europea de Dirección y Economía de la Empresa. 23,* 72-81.

Casas, M. (2016). *Acomodaciones computarizadas para la evaluación de la compresión lectora en estudiantes con limitación visual* (Tesis de maestría). Universidad Nacional de Colombia, Bogotá.

Chaffin, W. W. y Talley, W. K. (1980). Individual Stability in Delphi Studies. Technological Forecasting and Social Change 16, 67-73.

Cizek, G. J., Rosenberg, S. L., y Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*, 397-412.

Cizek, G. J., Bowen, D. y Church, K. (2010a). Sources of Validity Evidence for Educational and Psychological Test: A Follow-Up Study. *Educational and Psychological Measurement, 70*, 732-743.

Cizek, G. J. (2010b). Error of measurement: Validity and the place of consequences. *NCME Newsletter, 18*, 4–5.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*, 31-43.

Cizek, G. J. (2016a). Validating test scores meaning and defending test score use: different aims, different methods. A*ssessment in Education: Principles, Policy and Practice, 23*, 212-225.

Cizek, G. J. (2016b). Progress on validity: the glass half full, the work half done. A*ssessment in Education: Principles, Policy and Practice, 23*, 304-308.

Congreso de la República de Colombia (2009). *Ley 1324 de 2009.* Recuperado en octubre 10, 2012 de http://www.mineducacion.gov.co/1621/articles-210697archivopdfley1324.pdf

Crocker, L. (1997). Editorial: The great validity debate. *Educational Measurement: Issues and Practice, 16*, 4–4, 34.

Cronbach, L. G. (1949). Essentials of Psychological Testing. Nueva York: Harper & Brothers.

Cronbach, L. J. y Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cronbach, L. J. (1971). Test validation. En R.L. Thorndike (Ed.) *Educational Measurement (Second Edition)*. Washington DC. American Council on Education, (pp. 443-507)

Cronbach, L. J. (1980). Validity on parole: how can we go straight? En: W.B. Schrader (Ed.) *Measuring Achievement: Progress over Decade. Proceeding of the 1979 Educational Testing Service Invitational Conference*. San Francisco, CA. Jossey-Bass (pp. 99-108)

Cronbach, L. J. (1988). Five perspectives on validity argument. En: H. Wainer y H. I. Braun (Eds.), *Test validity*, (pp. 3-17). Nueva York: Routledge.

Cruz, M. y Martínez, M. C. (2012). Perfeccionamiento de un instrumento para la selección de expertos en las investigaciones educativas. *Revista Electrónica de Investigación Educativa. 14* (2), 167-179.

Cuevas, M. L. (2013). *Sesgo cultural en los ítems de las pruebas del examen saber 11° en Colombia.* (Tesis de Maestría) Universidad Nacional de Colombia.

Curenton, E.E. (1951). Validity. En: E.F. Lindquist (Ed.) *Educational Measurement*, (pp. 621-694). Washington, DC: American Council on Education.

Custer, R. L., Scarcella, J. A. y Stewart, B. R. (1999). The modified Delphi technique – a rotacional modification. *Journal of Vocational and Technical Education 15* (2), 50-58.

Dajani, J. S., Sincoff, M. Z. y Talley, W. K. (1979). Stability and agreement criteria for the termination of Delphi studies. *Technological Forecasting & Social Change*. *13* 83–90.

Dalkey, N. y Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Managment Science*, *9* 458-467

Dalkey, N. C. (1969). *The Delphi method: An experimental study of group opinion*. Document n.º RM-5888-PR. California: RAND Corporation.

Davidson, P., Merritt-Gray, M., Buchanan, J. & Noel, J. (1997). Voices from practice: mental health nurses identify research priorities. *Archives of Psychiatric Nursing XI, 6* 340–345.

Elmer, F., Seifert, I., Kreibich, H. y Thieken, A. (2010). A Delphi Method Expert survey to Derive Standards for Flood Data Collection. Risk Analysis, 30, 107- 124.

Elosua, P. (2003). Sobre la validez de los test. *Psicothema*, *15*, 315-321

Escalante, E. (2009). Perspectivas en el análisis cualitativo. *Theoria, 18*, 55-67.

Espinosa, A. (2014). *Evaluación objetiva de los procesos cognitivos involucrados en la comprensión de lectura* (Tesis de maestría). Universidad Nacional de Colombia, Bogotá.

Freeman, F. N. (1917). Review of "Manual of mental and physical test". *Psychological Bulletin, 14*, 105-6.

Gallie, W. B. (1956). Essentially contested concepts. *Proceedings of the Aristotelian Society, 56*, 167–198.

García, P. S. y Lazzari, L. L., (2000). La evaluación de la calidad en la universidad. *Cuadernos del CIMBAGE. 3*, 81-97.

Garrett, H. E. (1937). *Statistics in Psychology and Education*. Nueva York: Longmans, Green.

Goodman, C. M. (1987). The Delphi technique: A critique. *Journal of Advanced Nursing, 12*, 729–734.

Green, B., Jones, M., Hughes, D. & Williams, A. (1999) Applying the Delphi technique in a study of GPs information requirement. Health and Social Care in the Community 7(3), 198–205.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427-439.

Hasson, F.; Keeney, S. y McKenna, H. (2000). Research guidelines for the Delphi Survey technique. *Journal of Advance Nursing, 32,* 1008-1015.

Herrera, A. N. (2005). *Efecto del tamaño de muestra y razón de tamaños de muestra en la detección de funcionamiento diferencial de los ítems* (Tesis de doctorado). Universidad de Barcelona, España.

Herrera, A. N., Gómez, J. e Hidalgo, M. D. (2005). Detección de sesgo en los ítems mediante análisis de tablas de contingencia. *Avances en Medición, 3*, 29-52.

Herrera, A. N., Gómez, J. y Muñiz, J. (2007). Detección del funcionamiento diferencial de los ítems en el marco de la teoría de respuesta al ítem. *Avances en Medición*, 5, 27-46.

Herrera, A. N., Gómez, J., Quintero, C., Arias, E. M., Berrio, A. I. y Cervantes, V.H. (2007). *Identificación de ítems con sesgo cultural en las pruebas de los exámenes de Estado en Colombia*. Proyecto de investigación. Manuscrito no publicado. Universidad Nacional de Colombia, Bogotá.

Hsieh, C.-H., Tzeng, F.-M. Wu, C.-G., Kao, J.-S., y Lai, Y.-Y. (July-August 2011). The comparison of online Delphi and Real-Time Delphi. In Proceedings of the 2011 *Technology Management in the Energy Smart World (PICMET)*, Taoyuan, Taiwan.

Hubley, A. M. y Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research: An international interdisciplinary journal for quality of life measurement, 103,* 219-230.

Jeste, D., Ardelt, M., Blazer, D., Kraemer, H.C., Vaillant, G. y Meeks, T.W. (2010). Experts Consensus on Characteristics of Wisdom: A Delphi Method Study. *The Gerontologist, 50*, 668-680.

Jorm, A.F. (2015). Using the Delphi expert consensus method in mental health research. *Australian & New Zealand Journal of Psychiatry, 49* (10), 887-897.

Kauko, K. y Palmroos, P. (2014). The Delphi method in forecasting financial markets – An experimental study. *International Journal of Forecasting 30*, 313-327.

Kaplan, A., Skogstad, A. L. y Girshick, M. (1949). *The Prediction of Social Technological Events.* Document n.º P93, California: RAND Corporation.

Kane, M. (1992). An argument-based approach to validity, *Psychological Measurement: Issues and Practice, 21*, 31-41.

Kane, M. (2001). Current Concerns in Validity Teory. *Journal of Educational Measurement, 38*, 319-342.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Kane, M. (2016a). Explicating validity. A*ssessment in Education: Principles, Policy and Practice, 23*, 198-211.

Kane, M. (2016b). Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, Policy and Practice, 23*, 309-311.

Keeney, S., Hasson, F. y McKenna, H. P. (2001). A critical review of the Delphi technique as a research methodology for nursing. *International Journal of Nursing Studies, 38*, 195–200.

Keeney, S., Hasson, F. y McKenna, H. P. (2006). Consulting the oracle: Ten lessons from using the Delphi technique in nursing research. *Journal of Advanced Nursing, 53*(2), 205–212.

Keeney, S., Hasson, F. y McKenna, H. (2011). *The Delphi Technique in Nursing and Health Research*. Oxford: Wiley-Blackwell.

Kelly, T. L. (1927). *Interpretation of Educational Measurements*. Nueva York: World Book Company.

Landeta, J. (2006). Current validity of the Delphi method in social sciences. *Technological Forecasting & Social Change*. *73*, 467-482.

Landeta, J., Barrutia, J. y Lertxundi, A. (2011). Hybrid Delphi: A methodology to facilitate contribution from experts in professional contexts. *Technological Forecasting & Social Change*. *78*, 1629-1641.

Lancheros, L. C. (2013). *Métodos de equiparación de puntuaciones: los exámenes de*

estado en la población con y sin limitación visual (Tesis de maestría). Universidad Nacional de Colombia, Bogotá.

Lissitz, R. W. (2009). Introduction. In: R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications.* (pp. 1-15). United States of America: Information Age Publishing, INC.

Lissitz, R. W. y Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36,* 437-448.

Linstone, H.A. & Turoff, M. (1975) *The Delphi Method: Techniques and Applications.* Addison-Wesley, Reading, Massachusetts. Longhurst, R. (1971) An Economic Evaluation of Human.

Linstone, H. A. y Turoff, M. (Ed.) (2002). *The Delphi Method: Techniques and Applications*. Massachusetts: Addison-Wesley, Reading. Recuperado el 20 de marzo, 2014 de http://is.njit.edu/pubs/delphibook/. Trabajo original publicado en 1975.

Ma, Z. Shao, C. y Ye, Z. (2011). Constructing road safety performance indicators using Fuzzy Delphi Method and Grey Delphi Method. *Experts Systems with Applications, 38*, 1509-1514.

Marchais-Roubelat, A. y Roubelat, F. (2011). The Delphi method as a ritual: Inquiring the Delphi oracle. *Technological Forecasting and Social Change, 78,* 1491-1499.

Markus, K. A. y Borsboom, D. (2013). *Frontiers of test validity theory.* Nueva York: Routledge.

Markus, K. A. (2016). *Alternative vocabularies in the test validity.* A*ssessment in Education: Principles, Policy and Practice, 23,* 252-267.

Martinez, R. (1996). *Psicometría: Teoría de los tests psicológicos y educativos.* Madrid: Síntesis.

McKenna, H. P. (1994). The Delphi technique: A worthwhile research approach for nursing? *Journal of Advanced Nursing, 19,* 1221-1225.

Messick, S. y Anderson S. (1974). Educational testing, individual development and social responsibility. En R.W. Tyler y R.M. Wolf (Eds). *Crucial Issues in Testing,* Berkely, CA: Mc Cutchan Publishing Corporation. (pp. 21-34).

Messick, S. (1980). Test validity and the ethics of assessment. *American psychologist, 35*, 1012-1022.

Messick, S. (1988). The once and future issues of validity: assessing the meaning and consequences of measuring. En: H. Wainer y H. I. Braun (Eds.), *Test validity*, (pp. 33-45). Nueva York: Routledge.

Messick, S. (1989). Validity. En: R. L. Linn (Ed.), *Educational Measurement*. (pp. 13-103). Washington, D.C.: American Council on Education.

Meyer, M. (1908). The Grading of Students, *Science, 28*, 243-250.

Mitchell, V. W. (1991). The Delphi technique: An exposition and application. *Technology Analysis & Strategic Management, 3*(4), 333–358.

Michell, J. (1999). *Measurement in psychology.* Cambridge: Cambridge University Press.

Michell, J. (2009). Invalidity in validity. En: R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications.* (pp. 111-133). Estados Unidos: Information age Publishing, INC.

Monroe, W. S. (1923). *An introduction to the Theory of Educational Measurements*. Cambridge: Riverside Press.

Moss, P. A. (1998). The role of consequences in Validity Theory. *Educational Measurement: Issues and Practice, 17*, 6-12.

Moss, P. A., Girard, B. J. y Haniford, L.C. (2006). Validity in educational assessment. *Review of Research in Education, 30*, 109-162.

Moss, P. A. (2007). Reconstructing validity. *Educational Researcher, 36*, 470-476.

Moss, P. A. (2016). Shifting the focus of validity for test use. A*ssessment in Education: Principles, Policy and Practice, 23*, 236-251.

Muñiz, J. (1996). *Psicometría.* Madrid: Universitas S. A.

Murphy, M.K.; Sanderson, C.F.B.; Black, N.A; Askham, J.; Lamping, D.L.; Marteau, y McKee, C.M. (1998). Consensus development methods, and their use in clinical guideline development. *Health Technological Assessment, 2*, 5–83.

Newton, P. E. (2013a). Standards for talking and thinking about validity. *Psychological Methods, 18*(3), 301-319.

Newton, P. E. (February 2013b). Does it matter what 'validity' means? *Department of Education, Public Seminars.* The University of Oxford. Recuperado el marzo 15, 2015 de http://oucea.education.ox.ac.uk/wordpress/wp-content/uploads/2013/06/2013-Meaning-of-validity-Oxford-v4-slides.pdf

Newton, P. E. y Shaw, D. S. (2014). *Validity in educational and psychological assessment.* Londres: Cambridge Assessment.

Newton, P. E. y Shaw, D. S. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. A*ssessment in Education: Principles, Policy and Practice, 23,* 178-197.

NVivo (2016). (Versión pro 11.4.0) [software]. QSR International Pty. Ltda. Obtenido de: http://www.qsrinternational.com/product/nvivo-mac

Okoli, C. y Pawlowski, S. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information and Management, 42,* 15-29.

Ortega, F. (2008). El método Delphi, prospectiva en Ciencias Sociales a través del análisis de un caso práctico. *Revista EAN, 64,* 31-54.

Padilla, J.L. y Benitez, I. (2014). Validity evidence base don response processes. *Psicothema, 26,* 136-144.

Pankratova, N.D. y Malafeeva, L.Y. (2012). Formalizing the consistency of experts' judgments in the Delphi method. *Cybernetics and Systems Analysis, 48,* 711– 721.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, hereditary and panmixia. *Philosophical Transactions of the Royal Society A, 187,* 253-318.

Powell, C. (2003). The Delphi Technique: Myths and Realities. *Methodological Issues in Nursing Research, 41*(4), 376–382.

Rauch, W. (1979). The Decision Delphi. *Technological Forecasting and Social Change 15,* 159-169.

Reid, N. G. (1988). The Delphi technique, its contribution to the evaluation of professional practice. En: R. Ellis (Ed.). *Professional Competence and Quality Assurance in the Carring Professions.* (pp. 230-254), New York: Chapman and Hall.

Rico, J. D. (2015). *Análisis de Sesgo en las pruebas SABER 2009* (Tesis de maestría). Universidad Nacional de Colombia, Bogotá.

Rodriguez, D. (2016). *Evaluación empírica de dos métodos de equiparación. Búsqueda de una alternativa para garantizar la equidad de la medición de personas con limitación visual en pruebas masivas* (Tesis de maestría). Universidad Nacional de Colombia, Bogotá.

Ruch, G. M. (1924). *The improvement of the written examination*. Chicago: Scott, Foreman and Company.

Ruch, G. M. (1929). *The objective or New-type Examination: An introduction to educational measurement*. Chicago: Scott, Foresman and Company.

Ruch, G. M. (1933). Educational test and their uses: Recent developments in statistical prediction. *Review of Research in Education, 3,* 33-40.

Rulon, P. J. (1946). On the validity and the educational test. *Harvard Educational Review*, *16*, 290-296.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*, 5-18, 13, 24.

Shepard, L. A. (2016). Evaluating test validity: reprise and progress. *Assessment in Education: Principles, Policy & Practice, 23,* 268-280.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*, 477-481.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence. En: R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications.* (pp. 19-37). Estados Unidos: Information age Publishing, INC.

Sireci, S. G y Sukin, T. (2013). Test Validity. En: K. F. Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology. V1* (pp. 19-37). Estados Unidos: United Book Press. (pp. 61-84)

Sireci, S. y Padilla, J.L. (2014). Validating assessments: Introduction to the special section. *Psicothema, 26*, 97-99.

Sireci, S. G. (2016a). On the validity of useless test. *Assessment in Education: Principles, Policy & Practice, 23,* 226-235.

Sireci, S. G. (2016b). Comments on valid (and invalid?) commentaries. *Assessment in Education: Principles, Policy & Practice, 23,* 319-321.

Skulmoski, G. J., Hartman, F. T. y Krahn, J. (2007). The Delphi Method for Graduate Research. *Journal of Information Technology Education, 6*, 1-21.

Soler, M. (2014). *Evaluación de la comprensión de lectura en personas con limitación visual* (Tesis de maestría). Universidad Nacional de Colombia, Bogotá.

Spearman, C. (1904). General intelligence: Objectively determined and measured. *American Journal of Psychology, 18*, 161-169.

Starch, D. y Elliot, E. C. (1912). Reliability of the grading of high-school work in English. *The School Review*, *20*(7), 447-457.

Starch, D. y Elliot, E. C. (1913). 'Reliability of the grading work in Mathematics'. *The School Review*, *21*(4), 254-259.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(268), 677-680.

Suppes, P. y Zinnes, J. L. (1962). *Basic measurement theory*. *Technical Report Nº 45*. Stanford: Stanford University. Recuperado en octubre 18, 2012 de http://suppescorpus.stanford.edu/techreports/IMSSS_45.pdf

Thompson, J. B. (1990). Ideology and modern culture. In *The methodology of interpretation* (pp. 272-327). Stanford Stanford University Press.

Thorndike, E. L. (1903). *Educational Psychology*. Nueva York: Teachers College, Columbia University.

Thorndike, E. L. (1904). *An Introduction to the Theory of Mental and Social Measurements*. Nueva York: Teachers College, Columbia University.

Thurstone, L. L. (1932). *The reliability and validity of test*. Ann Arbor. MI: Edwards Brothers.

Von der Gracht, H. A. (2012). Consensus measurement in Delphi studies: Review and implications for future quality assurance. *Technological Forecasting & Social Change, 79*, 1525–1536.

Weir, A.; Hölmich, P.; Schahe, A. G. Delahunt, E. y de Vos, R. (2015). Terminology and definitions on grain pain in athletes: building agreement using a short Delphi method. *Journal of Sports Medicine, 49*, 825-827.

Wentholt, M.T.A., Rowe, G., König, A., Marvin, H.J.P., Frewer, L. J. (2009). The views of key stakeholders on an evolving food risk governance framework: Results from a Delphi study. *Food Policy, 34*, 539- 548.

Ziglio, E. (1996) The Delphi method and its contribution to decision-making. En: M. Alder & E. Ziglio (Eds.), *Gazing Into the Oracle: the Delphi Method and Its Application to Social Policy and Public Health*. (pp. 3–33). Jessica Kingsley Publishers, London.

Zumbo, B.D. y Hubley, A.M. (2016). Bringing consequences and side effects of testing and assessment to the foreground. *Assessment in Education: Principles, Policy & Practice, 23*, 299-303.

Apéndice A. Cuestionario n.º 1

Talking About the Concept of Validity!

General information

Thank you for participating in this study!

In this first survey, you will find two parts. The first one is about general contact information and the second one is related to the concept of Validity.

The six questions about the concept of Validity are open ended and ask for your own opinion about the concept. Please explain your opinion in a few paragraphs (no more than ten paragraphs) and submit your answers before January 22, 11:59 (-5GMT).

Remember that in this study:

- All participants are experts on validity.
- Participation is anonymous in as much there is no face to face interaction between the experts and their names are not revealed to the other experts during the study.
- Surveys follow a sequence given by the study researchers.
- Two or more rounds of questionnaires usually take place before arriving at some stable results.

Enjoy!

1. Please complete the following information.

First Name *

Last Name *

Institution *

Role *

Country *

Email Address *

Validity

2. What is validity in the context of educational or psychological measurement? *

3. What is/are the proper object(s) of validity? *

4. What is your opinion about the concept of validity described in the Standards for Educational and Psychological Testing, 2014 Edition (AERA, APA & NCME)? *

5. Which do you think are the difficulties that hinder achieving a consensus on the concept of validity? *

6. How can these difficulties be overcome? *

7. Based on your previous responses, how should a validation process be carried out? *

You have completed this survey!

We truly appreciate your collaboration with this project.

The next step will be to analyze the information, and based on it, we will make the next questionnaire. This new questionnaire will consist of rating scale items. The results of this round and the new instrument will be sent during the second week of February.

We hope to count on your valuable participation in the following rounds.

Sandra Camargo

Apéndice B. Cuestionario n.º 2

Agreement About the Concept of Validity!

Introduction

Welcome to the second questionnaire!

This new questionnaire consists of statements about validity derived from the opinions you and the other experts expressed on the first questionnaire. The aim of this questionnaire is to identify the degree of agreement you have with these ideas about the concept of validity and related issues.

This questionnaire has four sections that ask your opinion of some ideas about

- The concept of validity.
- Validity in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014).
- Problems in reaching consensus about the concept of validity, and whether a consensus is desirable.
- Validation.

You may save your progress during the completion of the survey by clicking on the ¨Save and continue later¨ button.
Please submit your answers before February 27, 11:59 (-5GMT).

Please rate your opinion on each statement by moving the bar from 0 (strongly disagree) to 10 (strongly agree). An example of the item style follows:

The best place to vacation is the beach. *

0- Strongly disagree — Neutral — 10- Strongly agree

Before to begin, please write your full name. *

About the Concept of Validity

1. Validity is the degree to which collected evidence, theory, and logical argument support the intended inferences to be made from test scores. *

0-Strongly disagree — Neutral — 10- Strongly agree

2. Validity is the degree to which it is possible to measure whatever needs to be measured by implementing the assessment procedure for the testing. *

0- Strongly disagree — Neutral — 10- Strongly agree

3. Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. *

0- Strongly disagree — Neutral — 10- Strongly agree

4. Validity is the extent to which the claims (interpretations and score-based decisions) based on assessment scores are adequately supported by evidence. *

0- Strongly disagree — Neutral — 10- Strongly agree

5. Validity designates the ability of a measurement instrument to detect variation between measured entities in an attribute of interest. *

0- Strongly disagree — Neutral — 10- Strongly agree

6. Validity is the extent to which the inferences that are made on the basis of the outcomes of the assessment are meaningful, useful and appropriate. *

0- Strongly disagree — Neutral — 10- Strongly agree

7. Validity is not intended to encompass ethical evaluation of test score use. *

0- Strongly disagree — Neutral — 10- Strongly agree

8. Validity is a socially situated process that encompasses the uses and consequences of the measurements evaluated. *

0- Strongly disagree — Neutral — 10- Strongly agree

9. The object of investigation in validity is the quality of the argumentation and evidentiary backing for a given interpretation and use of scores. *

0- Strongly disagree — Neutral — 10- Strongly agree

10. The proper object of validity is the interpretation or meaning of a score. *

0- Strongly disagree — Neutral — 10- Strongly agree

11. The proper object of validity is the measurement instrument. *

0- Strongly disagree — Neutral — 10- Strongly agree

12. The proper object of validity cannot be determined unless there is some agreement about what validity actually means. *

0- Strongly disagree — Neutral — 10- Strongly agree

13. The consequences of testing are not a source of validity evidence. *

0- Strongly disagree — Neutral — 10- Strongly agree

14. The consequences of testing are a source of evidence for justifying the use of a score. *

0- Strongly disagree — Neutral — 10- Strongly agree

15. The issue of the defensibility of test score use and the consequences of testing (ethical evaluation) is broader than validity. *

0- Strongly disagree — Neutral — 10- Strongly agree

16. The consequences of testing could be classified as part of the 'acceptability' of a testing procedure, where validity is nested within acceptability. *

0- Strongly disagree — Neutral — 10- Strongly agree

About the Standards

17. The Standards for Educational and Psychological Testing (2014) give a proper account of what validity is. *

0- Strongly disagree — Neutral — 10- Strongly agree

18. The Standards' definition of validity conflates the validity of score interpretation and the appropriateness of score use. *

0- Strongly disagree — Neutral — 10- Strongly agree

19. The consequences of testing should be removed from the Standards' definition of validity. *

0- Strongly disagree — Neutral — 10- Strongly agree

20. It is unclear in the Standards definition whether the proper object of validity is the intended interpretation of test scores or any number of possible actual interpretations drawn by any number of possible actual users. *

0- Strongly disagree — Neutral — 10- Strongly agree

21. The Standards (2014) focus on interpretations and do not give enough attention to how test scores are used, the contexts of use, and the consequences of use. *

0- Strongly disagree — Neutral — 10- Strongly agree

22. The Standards (2014) insufficiently recognize the importance of the basic methodological question of whether the test indeed measures the targeted attribute. *

0- Strongly disagree — Neutral — 10- Strongly agree

23. Given the definition of validity from the Standards, it is always possible to come up with support for a test score interpretation, especially if one is allowed to introduce any kind of evidence and propose any kind of score interpretation. *

0- Strongly disagree — Neutral — 10- Strongly agree

24. The continuing debate about the concept of validity is less important than a shared understanding of validation. *

0- Strongly disagree — Neutral — 10- Strongly agree

25. The importance of validity is widely enough recognized that it finds its way into laws and regulations. *

0- Strongly disagree — Neutral — 10- Strongly agree

26. The Standards' definition of validity is limited in the sense that it is ambiguous and can be interpreted in different ways. *

0- Strongly disagree — Neutral — 10- Strongly agree

27. The Standards (2014) are insufficient to guide validation practice. *

0- Strongly disagree — Neutral — 10- Strongly agree

28. Evidence based on "Test Development and Administration Procedures" should be added to the Standards. *

0- Strongly disagree — Neutral — 10- Strongly agree

29. The Standards are methodologically weak but politically strong. *

0- Strongly disagree — Neutral — 10- Strongly agree

30. Validity, as defined in the Standards in terms of "the degree to which evidence and theory support," is not restricted to educational and psychological measurement, and should be applicable in a wider context. *

0- Strongly disagree — Neutral — 10- Strongly agree

About Consensus

31. There are only small differences among the approaches to defining validity. *

0- Strongly disagree — Neutral — 10- Strongly agree

32. The Standards (2014) achieved consensus on defining validity. *

0- Strongly disagree — Neutral — 10- Strongly agree

33. The Standards' definition of validity is limited in the sense that respected scholars (and who-knows-how-many measurement specialists and practitioners) do not fully accept this definition. *

0- Strongly disagree — Neutral — 10- Strongly agree

34. Differences in views about validity primarily represent differences of opinion about how the field of educational measurement should function now and in the future. *

0- Strongly disagree — Neutral — 10- Strongly agree

35. A consensus definition of validity would help to improve communication among practitioners and stakeholders, as well as to improve validation practice. *

0- Strongly disagree — Neutral — 10- Strongly agree

36. If there is no consensus about the meaning of validity (whether by formal definition or by the way it is used), then effective communication is not possible. *

0- Strongly disagree — Neutral — 10- Strongly agree

37. When we do talk about validity – amongst ourselves, to our students, to our stakeholders – there is this slight air of ambiguity over what, exactly, it is that we are supposed to be talking about. *

0- Strongly disagree — Neutral — 10- Strongly agree

38. Because test use is intertwined not only with scientific research, but also with policy evaluation and societal political issues, accepting score interpretation or score use as the central question of validity has important consequences for what test users are supposed to do. *

0- Strongly disagree — Neutral — 10- Strongly agree

39. Advances in technology make it possible to carry out highly innovative forms of assessment. To not have a conception of validity that applies to these forms of assessment is to invite misinterpretations and unsound practices. *

0- Strongly disagree — Neutral — 10- Strongly agree

40. It is necessary to know what a test measures. *

0- Strongly disagree — Neutral — 10- Strongly agree

41. It is necessary to know the specific evidence backing the interpretation of test scores as measuring a particular attribute. *

0- Strongly disagree — Neutral — 10- Strongly agree

42. It is necessary to know whether the intended practical use of the test scores is justified. *

0- Strongly disagree — Neutral — 10- Strongly agree

43. A way to reach agreement on some definitional issues on validity would be to implement discussion strategies focusing definitional issues about validity specific points. *

0- Strongly disagree — Neutral — 10- Strongly agree

About Consensus

44. One problem in reaching a consensus over the definition of validity is that scholars approach different kinds of questions (ontological, methodological, ethical). *

0- Strongly disagree — Neutral — 10- Strongly agree

45. One problem in reaching a consensus over the definition of validity is that there are too many theorists with perspectives who don't have practical experience doing validation work. *

0- Strongly disagree — Neutral — 10- Strongly agree

46. One problem in reaching a consensus over the definition of validity is that philosophical orientations among both theorists and practitioners are too disparate. *

0- Strongly disagree — Neutral — 10- Strongly agree

47. One problem in reaching a consensus over the definition of validity is that very few measurement specialists have a solid understanding of the educational and psychological measurement literature on validity. *

0- Strongly disagree — Neutral — 10- Strongly agree

48. One problem in reaching a consensus over the definition of validity is that the few scholars who are actively engaged in the topic find themselves drawn into philosophical debates about the concept of validity that then make it harder for non-specialists to access the literature. *

0- Strongly disagree — Neutral — 10- Strongly agree

49. One problem in reaching a consensus over the definition of validity is that different stakeholders debate validity from different perspectives, and often want to 'solve' different problems with their preferred definition. *

0- Strongly disagree — Neutral — 10- Strongly agree

50. One problem in reaching a consensus over the definition of validity is that the view of policymakers and the public at large remains rooted in historical conceptions that pay less attention to cognitive and social factors. *

0- Strongly disagree — Neutral — 10- Strongly agree

51. The differences in approaches to validity cannot be overcome. We have to live with them. *

0- Strongly disagree — Neutral — 10- Strongly agree

52. The differences in approaches to validity should not be overcome. We should embrace a certain amount of pluralism. *

0- Strongly disagree — Neutral — 10- Strongly agree

53. The differences in approaches to validity should not be overcome. The debates about validity arise because it is a foundational concept. *

0- Strongly disagree — Neutral — 10- Strongly agree

54. A consensus may be reached by seeking clarity about the reasons why different scholars choose to promote different definitions; that is, not what they are trying to argue about validity but why, which is not always clear. *

0- Strongly disagree — Neutral — 10- Strongly agree

55. We should try to be explicit about how our views of validity follow from antecedent philosophical, methodological, and/or political convictions. *

0- Strongly disagree — Neutral — 10- Strongly agree

56. Any particular definition of validity will affect the definition of all sorts of other 'assessment quality' concepts (e.g. reliability, fairness). So we should think in terms of defining sets of 'quality concepts' within 'quality frameworks' rather than thinking simply in terms of defining a single 'quality concept', validity. *

0- Strongly disagree — Neutral — 10- Strongly agree

About Validation

57. Validity theory should provide a conceptual framework to guide validation practice. *

0- Strongly disagree — Neutral — 10- Strongly agree

58. Extensions of argument-based validation procedures, applied to new forms of assessment, will give examples. *

0- Strongly disagree — Neutral — 10- Strongly agree

59. Extensions of argument-based validation procedures, applied to new forms of assessment, will set expectations for sound use of assessments. *

0- Strongly disagree — Neutral — 10- Strongly agree

60. A theory of validity ought to help us understand the meaning of the concept (validity), what it refers to, how it relates to other concepts and, importantly, how to demonstrate it (validation). *

0- Strongly disagree — Neutral — 10- Strongly agree

61. If validity theory is confused or contested, then emerging frameworks will lack clarity in the practical guidance required for their execution. *

0- Strongly disagree — Neutral — 10- Strongly agree

62. One might adopt a fairly narrow definition of 'validity' while still adopting a far broader definition of 'validation'. *

0- Strongly disagree — Neutral — 10- Strongly agree

63. To improve validation practices, the explication of assessments, contexts, and evidentiary bases should be more comprehensive than it typically is in current practice. *

0- Strongly disagree — Neutral — 10- Strongly agree

64. Validation is achieved when accumulated empirical evidence is enough to show that the inferences based on test scores are reasonable. *

0- Strongly disagree — Neutral — 10- Strongly agree

65. Validation consists of gathering evidence that confirms (or potentially disconfirms) an intended score interpretation. *

0- Strongly disagree — Neutral — 10- Strongly agree

66. Validation should evaluate the intended test score uses by gathering evidence that supports (or potentially does not support) using the test scores for prespecified purposes. *

0- Strongly disagree — Neutral — 10- Strongly agree

67. Validation is achieved when one shows that the test evokes a causal process (e.g., an item response process) such that its outcome indeed depends on the value of the targeted attribute. *

0- Strongly disagree — Neutral — 10- Strongly agree

68. Validation should embrace anything related to the scientific evaluation of 'measurement quality' for a particular assessment procedure. *

0- Strongly disagree — Neutral — 10- Strongly agree

69. Validation would include all sorts of sources of empirical evidence and logical analysis, including evidence of the consequences of testing. *

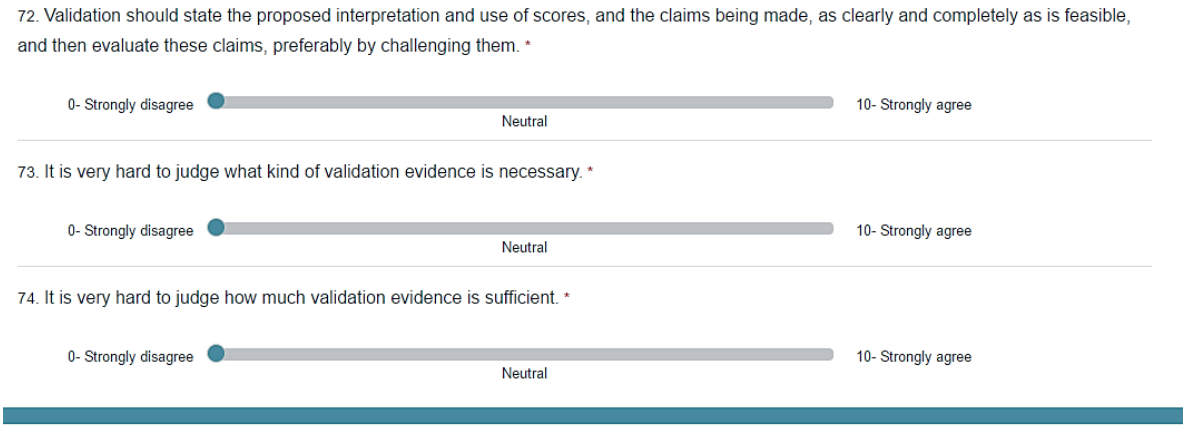
0- Strongly disagree — Neutral — 10- Strongly agree

70. During validation, the consequences of testing should be evaluated only with respect to how they affect the meaning of scores. *

0- Strongly disagree — Neutral — 10- Strongly agree

71. The social value of an assessment should be evaluated, but not as part of its validation. *

0- Strongly disagree — Neutral — 10- Strongly agree



You have completed this survey!

We truly appreciate your collaboration on this project.

The next step will be to analyze the information, and based on it, we will make a report of results, and design the next questionnaire. The report will show you both your own answers and the group's answers. This feedback may help you to have a panoramic view of the results, and provide a basis for response to the next survey.

The report of the results and the new instrument will be sent during March.
We hope to count on your valuable participation in the final round.

Sandra Camargo

Apéndice C. Cuestionario n.º 3

Agreement About the Concept of Validity - Final Part

Introduction

Welcome to the final questionnaire!

This questionnaire consists of statements about validity taken from the previous questionnaire, on which intra-group agreement was low (the interquartile range was relatively large). Please review the report you were sent via email, which contains both your own responses and summaries of the group's responses. We ask that you reconsider, again, each statement in the context of the group's responses

Then, please record your current opinion of each statement by moving the bar from 0 (strongly disagree) to 10 (strongly agree). You may change your original response, if you wish, but are not required to do so.

You may save your progress during the completion of the survey by clicking on the ¨Save and continue later¨ button.

Please submit your answers before April 7, 11:59 (-5GMT).

Before to begin, please write your full name. *

About the Concept of Validity

1. Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. *

0- Strongly disagree — Neutral — 10- Strongly agree

2. Validity is the extent to which the claims (interpretations and score-based decisions) based on assessment scores are adequately supported by evidence. *

0- Strongly disagree — Neutral — 10- Strongly agree

3. Validity designates the ability of a measurement instrument to detect variation between measured entities in an attribute of interest. *

0- Strongly disagree — Neutral — 10- Strongly agree

4. Validity is not intended to encompass ethical evaluation of test score use. *

0- Strongly disagree — Neutral — 10- Strongly agree

5. Validity is a socially situated process that encompasses the uses and consequences of the measurements evaluated. *

0- Strongly disagree — Neutral — 10- Strongly agree

6. The object of investigation in validity is the quality of the argumentation and evidentiary backing for a given interpretation and use of scores. *

0- Strongly disagree — Neutral — 10- Strongly agree

7. The proper object of validity is the interpretation or meaning of a score. *

0- Strongly disagree — Neutral — 10- Strongly agree

8. The proper object of validity is the measurement instrument. *

0- Strongly disagree — Neutral — 10- Strongly agree

9. The consequences of testing are not a source of validity evidence. *

0- Strongly disagree — Neutral — 10- Strongly agree

10. The issue of the defensibility of test score use and the consequences of testing (ethical evaluation) is broader than validity. *

0- Strongly disagree — Neutral — 10- Strongly agree

11. The consequences of testing could be classified as part of the 'acceptability' of a testing procedure, where validity is nested within acceptability. *

0- Strongly disagree — Neutral — 10- Strongly agree

About the Standards

12. The Standards for Educational and Psychological Testing (2014) give a proper account of what validity is. *

0- Strongly disagree — Neutral — 10- Strongly agree

13. The Standards' definition of validity conflates the validity of score interpretation and the appropriateness of score use. *

0- Strongly disagree — Neutral — 10- Strongly agree

14. The consequences of testing should be removed from the Standards' definition of validity. *

0- Strongly disagree — Neutral — 10- Strongly agree

15. The Standards (2014) insufficiently recognize the importance of the basic methodological question of whether the test indeed measures the targeted attribute. *

0- Strongly disagree — Neutral — 10- Strongly agree

16. The continuing debate about the concept of validity is less important than a shared understanding of validation. *

0- Strongly disagree — Neutral — 10- Strongly agree

17. The importance of validity is widely enough recognized that it finds its way into laws and regulations. *

0- Strongly disagree — Neutral — 10- Strongly agree

18. The Standards are methodologically weak but politically strong. *

0- Strongly disagree — Neutral — 10- Strongly agree

About Consensus

19. The Standards (2014) achieved consensus on defining validity. *

0- Strongly disagree — Neutral — 10- Strongly agree

20. The Standards' definition of validity is limited in the sense that respected scholars (and who-knows-how-many measurement specialists and practitioners) do not fully accept this definition. *

0- Strongly disagree — Neutral — 10- Strongly agree

21. Differences in views about validity primarily represent differences of opinion about how the field of educational measurement should function now and in the future. *

0- Strongly disagree — Neutral — 10- Strongly agree

22. If there is no consensus about the meaning of validity (whether by formal definition or by the way it is used), then effective communication is not possible. *

0- Strongly disagree — Neutral — 10- Strongly agree

23. Advances in technology make it possible to carry out highly innovative forms of assessment. To not have a conception of validity that applies to these forms of assessment is to invite misinterpretations and unsound practices. *

0- Strongly disagree — Neutral — 10- Strongly agree

24. It is necessary to know what a test measures. *

0- Strongly disagree — Neutral — 10- Strongly agree

25. It is necessary to know whether the intended practical use of the test scores is justified. *

0- Strongly disagree — Neutral — 10- Strongly agree

26. A way to reach agreement on some definitional issues on validity would be to implement discussion strategies focusing on specific points. *

0- Strongly disagree — Neutral — 10- Strongly agree

27. One problem in reaching a consensus over the definition of validity is that there are too many theorists with perspectives who don't have practical experience doing validation work. *

0- Strongly disagree — Neutral — 10- Strongly agree

28. The differences in approaches to validity should not be overcome. We should embrace a certain amount of pluralism. *

0- Strongly disagree — Neutral — 10- Strongly agree

29. The differences in approaches to validity should not be overcome. The debates about validity arise because it is a foundational concept. *

0- Strongly disagree — Neutral — 10- Strongly agree

30. Any particular definition of validity will affect the definition of all sorts of other 'assessment quality' concepts (e.g. reliability, fairness). So we should think in terms of defining sets of 'quality concepts' within 'quality frameworks' rather than thinking simply in terms of defining a single 'quality concept', validity. *

0- Strongly disagree — Neutral — 10- Strongly agree

About Validation

31. If validity theory is confused or contested, then emerging frameworks will lack clarity in the practical guidance required for their execution. *

0- Strongly disagree — Neutral — 10- Strongly agree

32. One might adopt a fairly narrow definition of 'validity' while still adopting a far broader definition of 'validation'. *

0- Strongly disagree — Neutral — 10- Strongly agree

33. Validation should embrace anything related to the scientific evaluation of 'measurement quality' for a particular assessment procedure. *

0- Strongly disagree — Neutral — 10- Strongly agree

34. Validation would include all sorts of sources of empirical evidence and logical analysis, including evidence of the consequences of testing. *

0- Strongly disagree — Neutral — 10- Strongly agree

35. During validation, the consequences of testing should be evaluated only with respect to how they affect the meaning of scores. *

0- Strongly disagree — Neutral — 10- Strongly agree

36. The social value of an assessment should be evaluated, but not as part of its validation. *

0- Strongly disagree — Neutral — 10- Strongly agree

37. It is very hard to judge what kind of validation evidence is necessary. *

0- Strongly disagree — Neutral — 10- Strongly agree

38. It is very hard to judge how much validation evidence is sufficient. *

0- Strongly disagree — Neutral — 10- Strongly agree